Catarina Alexandra Lopes do Vale

MODELAÇÃO E ESTIMAÇÃO DO RISCO DE CRÉDITO

ESTUDO DE UMA CARTEIRA

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Matemática e Aplicações - Actuariado, Estatística e Investigação Operacional.

Lisboa
2010

Catarina Alexandra Lopes do Vale

MODELAÇÃO E ESTIMAÇÃO DO RISCO DE CRÉDITO

ESTUDO DE UMA CARTEIRA

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Matemática e Aplicações - Actuariado, Estatística e Investigação Operacional.

Dissertação Orientada por:

Professora Doutora Gracinda Rita Diogo Guerreiro
Professor Doutor Manuel Leote Tavares Inglês Esquível

Lisboa
2010

Agradecimentos

Agradeço à minha família por se dedicar com tanto afinco à minha formação, pelo seu apoio incessante e incentivo, pelo carinho, pela paciência nos dias menos fáceis e pela compreensão pelas minhas tão longas ausências.

Ao João, por toda a compreensão e estímulo transmitido, pelo amor, pela amizade e pelos intermináveis dias em que me ausentei.

Aos professores da FCT, pelos valiosos conhecimentos que me transmitiram e um especial agradecimento aos meus supervisores Professor Doutor Manuel Leote Esquível e Professora Doutora Gracinda Rita Guerreiro pelo seu apoio e encorajamento, pela experiência, profissionalismo e dedicação, pela disponibilidade, amizade e confiança em mim.

Por último, mas certamente não menos importante, aos meus amigos pela amizade, incentivo constante, companheirismo e apoio durante todos os momentos da minha vida e em especial durante este percurso.

Obrigado!

Resumo

Os modelos de Credit Scoring são modelos quantitativos utilizados frequentemente nas instituições financeiras com o intuito de medir e prever o risco de crédito, possuindo uma avultada importância no processo de concessão de crédito.

O presente trabalho centra-se no estudo de uma carteira de crédito à Habitação.

O objectivo desta dissertação é determinar a probabilidade de *default* de um cliente aquando de um novo pedido de crédito, com base em informações facultadas pelo mesmo.

Os dados da carteira de crédito foram utilizados para desenvolver o modelo de *Credit Scoring* de aprovação de crédito, através de uma técnica estatística multivariada, designadamente a Regressão Logística. O modelo obtido agrega variáveis como a idade do cliente, o estado civil, o valor do crédito, o número de prestações e o rendimento líquido. Algumas destas variáveis representam atributos que contribuem para o aumento da incapacidade de cumprimento das obrigações de crédito, enquanto que outras contribuem para a diminuição dessa incapacidade.

Para o modelo de aprovação de crédito, é essencial a estimação da probabilidade de *default*.

Mostra-se que o cálculo do *spread* mínimo que a instituição financeira deve aplicar em cada pedido de crédito é função da probabilidade de *default* e da taxa de recuperação. Um estudo equivalente ao da probabilidade de *default* poderá ser efectuado para estimar a taxa de recuperação.

Por fim, faz-se uma breve avaliação do modelo de aprovação de crédito obtido, e conclui-se que o modelo de regressão logística aplicado, tem capacidade de, a partir dos dados da carteira de crédito, recuperar a informação essencial à estimação da probabilidade de *default*.

Palavras chave: *Risco de Crédito, Modelos de Credit Scoring, Probabilidade de Default, Taxa de Recuperação, Spread, Regressão Logística, Cliente Incumpridor.*

Abstract

Credit Scoring models are quantitative models often used in financial institutions in order to measure and predict the credit risk, having a bloated importance in the process of granting credit.

This paper focuses on the study of a simulated housing loan portfolio.

The purpose of this dissertation is to determine the default probability of a client when makes a new request for credit, based on information provided by it.

Data from the loan portfolio is used to develop a Credit Scoring model for credit approval, through a multivariate statistical technique, such as logistic regression. The model contains variables such as client's age, marital status, the value of the credit amount, the number of payments and net income. Some of these variables represent attributes that contribute to the increased inability to fulfill the obligations of credit while others contribute to the decrease of that incapacity.

For the model of credit approval, it is essential to estimate the default probability .

It is shown that the minimum spread that the financial institution must apply to each credit application is a function of the default probability and recovery rate. An equivalent study could be performed to estimate the recovery rate.

Finally, we provide a brief evaluation of the credit approval model obtained, and it is concluded that the logistic regression model, is able to, according to portfolio data, retrieve the essential information for estimating the default probability.

Keywords: *Credit Risk, Credit Scoring Models, Probability of Default, Recovery Rate, Spread, Logistic Regression, Defaulting Client.*

Conteúdo

Tabela de Notações

NOTAÇÃO	*DESCRIÇÃO*
t	Instante no tempo.
τ	Momento em que ocorre o *default*.
$N(t)$	Indicador do processo de *Default*.
EAD	*Exposure at Default*.
$E(t)$	Montante que o credor perde, caso o devedor entre em *default* no tempo t.
$R(t)$ ou R	Taxa de Recuperação (*Recovery Rate*).
$L(t)$	*Loss Given Default-LGD*.
T	Maturidade.
$D(T)$	Default Loss.
$PD_i(t,T)$	Probabilidade de *default* de um devedor i deste t até T.
$PS(t,T)$	Probabilidade de sobrevivência desde t até T.
N	Nº de *defaults*, num certo intervalo de tempo.
λ	Probabilidade de *default* por unidade de tempo (1 ano).
$\mathbb{E}[D_i(T)]$	Perda Esperada Individual.
UEL	Perda inesperada de uma carteira *Unexpected Default Loss*.
r	Taxa de juro.
s	*Spread*.
$\mathbb{P}$	*Medida de probabilidade Natural*.
p	Probabilidade de *default* na medida natural.
$\mathbb{Q}$	*Medida de probabilidade Martingala*.
q	Probabilidade de *default* na medida martingala.
$X_t^{\boldsymbol{\theta}}$	Processo Estocástico.
$(X_t)_{t\geq 0}$	Variável que descreve a evolução do preço de um activo - Processo de Markov.
$(B_t)_{t\geq 0}$	Processo Browniano.
$B(t,T)$	Preço em t, de uma *default-free zero coupon bond* com vencimento em T.
$\tilde{B}(t,T)$	Preço em t, de uma *defaultable zero coupon bond*) com vencimento em T.
EDE	Equação Diferencial Estocástica.
INE	Instituto Nacional de Estatística.

Prefácio

O estudo e o desenvolvimento de modelos de *Credit Scoring* tornou-se uma tarefa fundamental para as instituições financeiras que necessitam de controlar o risco de crédito. Com as mudanças da economia mundial nas últimas décadas, ocasionadas essencialmente pela globalização, tornou-se necessário reestruturar os métodos de avaliação de risco. São várias as razões que ocasionaram uma necessidade urgente de gerir o risco de crédito de modo eficaz, nomeadamente as alterações macro-económicas dos países, o aumento da competição bancária, a expansão dos mercados de capitais, as alterações constantes da taxa de juro ou as margens de *spreads* mais competitivas e os valores voláteis das garantias reais.

Perante esta realidade, as instituições financeiras têm vindo a desenvolver modelos internos de gestão de risco de crédito com o intuito de oferecer ferramentas mais eficientes para a valorização da carteira, medição de riscos, apreçamento de novos empréstimos, potencializando os ganhos dos capitais emprestados adequando-os ao montante de capital que estes devem manter como parte da sua estrutura de capital.

Numa notícia do Jornal de Negócios, [Guerreiro, 2010] argumenta:

"A crise financeira foi um invento dos bancos anglo-saxónicos mas os bancos europeus foram bons aprendizes dos erros alheios. Um dos erros clamorosos foi o da prática dos *spreads* zero vírgula qualquer coisa, uma negação do risco que está hoje a sair cara.

O *spread* que se soma às taxas *Euribor* é simultaneamente um medidor de risco e o lucro do banco. Ao eliminá-lo quase até ao zero, os bancos estavam, portanto, a abdicar de medir risco no seu cliente e da sua margem de lucro. No início, muitos bancos davam com uma mão (a taxa baixa) o que tiravam com a outra (as comissões altas). Mas como os governos lhes foram proibindo as comissões abusivas, as cláusulas leoninas, a impossibilidade de transferência de contratos, essa face oculta do empréstimo foi sendo extinta.

Foi um comportamento predatório contra si mesmos: hoje, grande parte do balanço dos bancos comerciais está amarrado a créditos à habitação de dezenas de anos com "spreads" baixos - e nada pode fazer para contrariá-lo. Os contratos estão assinados, os *direitos adquiridos* são dos clientes.

Excepto os clientes que precisem de alterar condições contratuais do seu crédito. Uma necessidade do cliente é um pretexto para o banco: o *spread* sobe. O *spread* sobe? Então o cliente não altera nada. Certo?

Errado: há clientes que não têm alternativa. A reportagem de hoje do Negócios mostra casos de desemprego que obrigam a dilatar o prazo do empréstimo. E de divórcios em que o contrato tem de passar de dois para um titular. Neste caso, a taxa de esforço (parte do rendimento afecta ao pagamento das prestações) pode até continuar a ser cumprida, mas mesmo assim a taxa sobe. A oportunidade torna-se oportunismo."

A contradição é evidente, é aplicado um *spread* elevado para clientes com mais dificuldades.

Dito isto, o propósito desta dissertação será estudar estatisticamente um conjunto de características, que influenciam ou não, a ocorrência de incumprimento (*default*). O resultado deste estudo é um modelo constituído pelas características que melhor explicam a ocorrência de *default*, medindo a probabilidade de *default* de cada cliente.

No Capítulo 1 são apresentados conceitos, medidas e modelos de risco de crédito para clientes individuais e para carteiras de crédito de uma instituição financeira com o intuito de fazer um enquadramento geral dos modelos e técnicas utilizadas para modelação de risco de crédito.

No Capítulo 2 são desenvolvidas expressões que determinam o *spread* adequado a cada cliente, em função da *probabilidade de default* e da *taxa de recuperação* estimadas, sendo descritas várias abordagens que conduzem a resultados comparáveis.

O Capítulo 3, no qual se insere a parte prática deste trabalho, consiste na simulação de uma carteira de crédito à Habitação e no estudo da mesma através dos Modelos Lineares Generalizados, em particular através da Regressão Logística. Foi necessária uma pesquisa prévia sobre a informação que geralmente é pedida a um solicitante de crédito, com o objectivo de se simular a carteira de crédito e, através de métodos de selecção de variáveis, encontrar o modelo que melhor explica a ocorrência de *default*. Deste modo, através deste último modelo, é possível estimar a probabilidade de *default* de um novo solicitante de crédito.

Ficam em aberto algumas questões, que não foram analisadas neste trabalho, mas que podem ser objecto de estudos futuros:

- Estimação da taxa de recuperação, necessária para na determinação do *spread* mínimo a aplicar, mas que se deduz ser estimada de forma análoga à probabilidade de *default* recorrendo, no entanto, a variáveis explicativas distintas.

- Aplicação do modelo Vórtices Estocásticos, ver [Guerreiro e Mexia, 2008], a uma carteira de crédito dinâmica, isto é, agrupar os clientes em "classes de risco" (com base na probabilidade de *default* estimada), estimar as probabilidades de transição entre classes de risco e, assumindo uma carteira de crédito aberta (permitindo a subscrição de novos contratos e a anulação de contratos existentes) analisar, numa perspectiva temporal, a evolução da dimensão de cada uma das classes de risco, estimando o peso relativo de cada uma das classes dentro da carteira de crédito.

Neste trabalho, optou-se por utilizar alguns termos técnicos em inglês pois ainda não existe um consenso para a sua tradução. Embora seja feita uma tradução dos termos, no desenvolvimento deste trabalho utiliza-se, para quase todos, a sua versão original.

No final deste trabalho consta um pequeno glossário com alguns dos termos utilizados.

Lista de Tabelas

Lista de Figuras

Capítulo 1

Análise de Risco de Crédito

Neste Capítulo são apresentados conceitos, medidas e modelos de risco de crédito para clientes individuais e para carteiras de crédito de uma instituição financeira.

1.1 Modelos de Análise de Crédito - Generalidades

1.1.1 Modelos de Risco de Crédito

Modelo é, por definição, uma representação simplificada de algo real. Desta forma, algoritmos, fórmulas, sistemas ou regras que visam representar processos ou atributos reais relacionados com risco de crédito podem ser considerados modelos de risco de crédito.

Devido às suas características, os modelos facilitam a compreensão de um fenómeno e, eventualmente, a sua exploração. Os modelos de risco de crédito não são excepção.

Para [Caouette et al., 1998] a construção de um modelo de risco de crédito exige, em primeiro lugar, a identificação das variáveis que podem provocar a ocorrência de incumprimento. Segue-se a utilização de um conjunto de ferramentas para construir um modelo formal, com base num conjunto de dados que representem a carteira de crédito. Finalmente, devem ser aplicados testes para determinar se o modelo tem o desempenho esperado.

Então, pode dizer-se que o termo *modelo de risco de crédito* é uma denominação genérica que, para a sua construção, necessita de diversas ferramentas. Destas ferramentas, convém destacar as técnicas econométricas, como a análise *Logit* e análise *Probit*, redes neuronais, modelos de optimização, sistemas especialistas ou baseados em regras, sistemas híbridos que utilizam a computação, estimativa e simulação directas.

Segundo [Caouette et al., 1998], a finalidade dos modelos de risco de crédito é diversa. A aprovação de crédito, o apreçamento de crédito, a estratégia de cobrança, a determinação de *ratings* de crédito, são alguns dos objectivos visados pelos modelos de risco de crédito. Estes modelos têm ainda o intuito de analisar e mensurar o risco da ocorrência de *cash-flows*, determinar o valor adequado para um empréstimo ou um título. Em particular, estes modelos facilitam a tomada de decisão do avalista pois geram informação que não estaria disponível de outra forma ou que teria custos muito elevados.

Na literatura é possível encontrar referências a diferentes formas de classificação para os modelos de risco de crédito. Segundo [Saunders e Allen, 2002], os modelos de risco de crédito estão divididos em *sistemas especialistas*, modelos de *Credit Scoring*, modelos de *Credit Rating* e modelos de *portfólio*. Os três primeiros, no seu todo, formam um grupo designado Modelos de Classificação de Risco (abordagem teórica tradicional) e o último é baseado na teoria de carteiras, pertencente a uma abordagem mais actual de modelos de risco de crédito.

1.1.2 Crédito e Default

A palavra crédito significa confiança/segurança: crédito = crer, do latim *creditum*, *credere*.

Crédito é também a capacidade de assumir compromissos, quer para financiamentos quer para empréstimos, junto do sistema financeiro. Essa capacidade é analisada com base em parâmetros técnicos.

A cedência de crédito é uma relação contratual mediante a qual uma das partes, a credora, entrega um determinado montante a outra parte, a devedora, sob a promessa de pagamento do valor em dívida.

O credor é toda a pessoa titular de um crédito, ou, que tem a receber de outrem uma certa importância em dinheiro. Protegido pela lei, o credor possui a faculdade de exigir do devedor o cumprimento da obrigação ou o pagamento do crédito, quando este se torna exigível, isto é, se o contrato não é cumprido pelo devedor.

O risco da promessa de pagamento não ser cumprida por parte do devedor, designa-se risco de crédito. Pode dizer-se que o risco de crédito surge implicitamente no acto de emprestar uma quantia a alguém, uma vez que existe sempre a incerteza em relação ao retorno dessa quantia.

Default é a incapacidade para cumprir as condições de uma obrigação ou acordo, ou seja, é não fazer um pagamento em dívida. A palavra *Default* pode ser simplesmente, *Incumprimento*.

1.1.3 Análise de Crédito

Segundo [Araújo, 2006], a concessão de crédito constitui uma das principais actividades das instituições financeiras sendo a principal causa de insolvência das mesmas e, uma vez que o risco de crédito é inerente a qualquer operação financeira, é essencial que estas instituições procedam a uma análise e gestão rigorosa dos créditos que concedem.

Uma boa gestão do risco de crédito por parte das instituições financeiras, é indispensável para que se evite a insolvência das mesmas. A análise de crédito é um processo que deve reunir informações a respeito do tomador de crédito, com o intuito de avaliar a sua capacidade de cumprir com as suas obrigações e definir quanto à concessão ou não do crédito.

A análise de crédito pode ser tratada tendo em conta duas metodologias: a qualitativa e a quantitativa. A análise qualitativa remete para julgamentos subjectivos por parte do analista de crédito, em relação à capacidade de pagamento do tomador de crédito. Nesta abordagem, pessoas especializadas são encarregues de tomar a decisão sobre a concessão de crédito, utilizando critérios qualitativos e subjectivos. A análise quantitativa utiliza informações provenientes de modelos estatísticos e econométricos. Nesta última abordagem consideram-se os modelos de *Credit Scoring* e os modelos baseados na teoria de carteiras.

A principal vantagem da abordagem qualitativa é a especificidade com que é tratado cada caso. A principal desvantagem é a sua dependência na experiência do avaliador, o baixo volume de produção e o envolvimento pessoal do concedente de crédito.

As regras bem definidas em relação às características dos clientes e às operações de crédito, baseadas, em geral, em modelos estatísticos, são o factor positivo da análise quantitativa.

A análise de crédito clássica e mais antiga depende essencialmente da opinião de especialistas bem treinados, no entanto, actualmente, a necessidade de obter uma avaliação de risco mais aprofundada e mais correcta obrigou ao desenvolvimento de técnicas estatísticas de modo a garantir uma diminuição do risco de crédito das instituições financeiras. As técnicas de pesquisa estatística, tais como a análise de sobrevivência, redes neuronais, programação matemática e simulação probabilística contribuíram para o avanço das técnicas de mensuração do risco de crédito.

1.1.4 Sistemas Especialistas de Análise Subjectiva

Os sistemas especialistas são modelos cuja função é apoiar a análise de crédito clássica referida anteriormente. Estes modelos baseiam-se no julgamento subjectivo de profissionais especializados na área, cuja decisão de conceder ou não crédito é apenas da sua responsabilidade. Segundo [Saunders e Allen, 2002], o conhecimento especializado do avalista, o seu julgamento subjectivo, e a atribuição de pesos a certos factores-chave são implicitamente determinantes na decisão da concessão de crédito.

Existe uma grande variedade de factores que podem ser analisados, no entanto, a análise mais comum rege-se pelas principais características do proponente de crédito. Esta análise é guiada pela avaliação dos C's do crédito: Carácter, Capacidade, Capital, Colateral, Condições e Conglomerado. Após a análise das informações recolhidas relativas a estes factores, o analista pondera-os e com o auxílio de critérios subjectivos obtendo uma pontuação que o permite decidir. [Saunders e Allen, 2002] referem dois problemas para este tipo de análise: consistência e subjectividade.

Os C's do crédito

Os 5 $C's$ do crédito foram enunciados, em 1972, por Weston e Brigham. [Silva, 2000] acrescenta aos 5 $C's$ um sexto C, que se refere ao factor "Conglomerado", que actualmente é considerado nas análises de crédito das instituições financeiras.

1. *Carácter*: Informações referentes à índole, idoneidade e reputação do cliente;
2. *Capacidade*: Este C deve fornecer informações que possibilitem avaliar se as receitas e despesas do cliente permitem o cumprimento das obrigações a serem assumidas;
3. *Capital*: Informações referentes à estrutura de capital, endividamento, liquidez, lucro e outros índices financeiros obtidos por meio dos demonstrativos financeiros do cliente;
4. *Colateral*: É a capacidade da pessoa individual ou colectiva para oferecer garantias ao empréstimo;
5. *Condições*: Informações referentes à capacidade do cliente ou dos administradores empresariais de se adaptarem a situações conjunturais, ter agilidade e flexibilidade de se adaptar e criar mecanismos de defesa;
6. *Conglomerado*: Informações sobre a situação de outras empresas situadas no mesmo grupo económico e como poderão afectar a empresa em estudo.

1.1.5 Modelos de *Credit Scoring*

Os modelos de *Credit Scoring* permitem, recorrendo a técnicas estatísticas, estabelecer um processo de atribuição de pontuações às variáveis de decisão de credito. Este processo permite estimar a probabilidade de um dado cliente ser um cliente cumpridor ou incumpridor.

Estes modelos visam o isolamento de características que permitem distinguir os bons créditos dos maus créditos.

Os modelos de *Credit Scoring* são sistemas que os credores utilizam para determinar a sua decisão quanto à concessão ou não de um crédito a um cliente.

Para [Caouette et al., 1998], os modelos tradicionais de *Credit Scoring* atribuem pesos, determinados estatisticamente, a alguns atributos do solicitante, de modo a criar um *score* de crédito.

Segundo [Saunders e Allen, 2002] os sistemas de pontuação de crédito encontram-se em quase todos os tipos de análise de crédito. O objectivo é geralmente o mesmo; pré-identificar factores-chave que determinem a probabilidade de *default*, e a combinação ou ponderação dos mesmos de modo a produzir uma pontuação quantitativa.

A metodologia básica para o desenvolvimento de um modelo de *Credit Scoring* deve ter em conta, segundo [Saunders e Allen, 2002], as seguintes etapas:

- *Planeamento e definições*: mercados e produtos de crédito para os quais o sistema será desenvolvido; finalidades de uso; tipos de clientes; conceito de incumprimento a ser adoptado; horizonte de previsão do modelo;

- *Identificação das variáveis*: caracterização do proponente ao crédito; caracterização da operação; selecção das variáveis significativas para o modelo; análise das restrições a serem consideradas em relação às variáveis;

- *Planeamento amostral e colecta de dados*: selecção e dimensionamento da amostra; colecta dos dados; montagem da base de dados;

- *Determinação da fórmula de classificação através de técnicas estatísticas*: por exemplo, a análise discriminante ou a regressão logística;

- *Determinação do ponto de corte*, a partir do qual o cliente é classificado como cumpridor ou bom pagador; em outras palavras, é o ponto a partir do qual a instituição financeira pode aprovar a concessão do crédito;

A diferença mais acentuada entre os modelos subjectivos e o modelo de *Credit Scoring* reside no facto de, nesse último, se valorizar a utilização de métodos estatísticos que permitem a selecção dos factores-chave e dos respectivos pesos, a partir dos quais se obtém uma pontuação para cada cliente de acordo com as suas características. A pontuação gerada, fornece indicadores quantitativos das hipóteses de incumprimento desse cliente e representa o risco de perda. Esta pontuação pode ser comparada com um ponto de corte ou com uma pontuação mínima aceitável a partir da qual a instituição financeira aprova ou não a concessão de crédito.

[Lewis, 1992] refere que o primeiro modelo estatístico de análise de crédito remonta ao ano de 1945. Os primeiros modelos de *Credit Scoring* destinavam-se ao crédito ao consumo. A expansão do mercado de crédito massificado obrigou os analistas a uma maior rapidez e homogeneidade no tratamento dos seus clientes e, por isso, a um aumento da utilização destes modelos. A evolução dos sistemas informáticos possibilitou o tratamento estatístico adequado a esse aumento de dados.

Ainda que o uso de métodos de *Credit Scoring* seja essencialmente direccionado para a tomada de decisões sobre a concessão ou não de um crédito, algumas instituições também os usam para determinar o montante do crédito a ser concedido, como refere [Caouette et al., 1998].

As classificações destes modelos são classificações numéricas ou quantitativas. A pontuação de crédito é um número que resulta de um estudo baseado no histórico de pagamentos, no perfil de dívida do devedor e nos dados estatísticos de outros devedores que os credores utilizam para determinar a probabilidade de ocorrerem certos comportamentos de crédito, como o de um devedor falhar com as suas obrigações. A pontuação de crédito tem um impacto importante no *spread* que será atribuído quando se concede o crédito.

Mais recentemente, o *Credit Scoring* serve de base ao desenvolvimento de modelos mais complexos baseados na teoria de carteiras.

Nos modelos de *Credit Scoring* supõe-se que as características dos clientes que entrarão em *default* no futuro são semelhantes às características dos clientes que entraram em *default* no passado. Tendo em conta este pressuposto, é comum utilizar-se amostras de clientes cumpridores e incumpridores da instituição bancária e aplicar técnicas estatísticas apropriadas para inferir sobre os indícios de incumprimento de um cliente em particular.

Entre as variáveis de decisão de crédito escolhidas, algumas contribuem para o aumento outras para a diminuição do nível de incumprimento.

No âmbito da aplicação do modelo de *Credit Scoring* para a avaliação de risco de crédito e insolvência de empresas, as variáveis mais utilizadas são os índices de contabilidade e financeiros como a rentabilidade, a liquidez e o endividamento.

[Caouette et al., 1998] refere que o modelo *Z-Score* é um exemplo clássico de modelos de previsão de insolvência de empresas.

Z-*Score*

Com base em [Wikipédia, 2009d], sabe-se que a fórmula *Z-Score* para prever a falência foi publicada em 1968 por Edward I. Altman, que era, na época, um professor adjunto de Finanças na Universidade de Nova York. A fórmula pode ser usada para prever a probabilidade de uma empresa ir à falência no prazo de dois anos.

O *Z-Score* utiliza o rendimento das sociedades múltiplas e valores do balanço patrimonial para medir a saúde financeira de uma empresa.

O Z-*Score* é uma combinação linear de quatro ou cinco rácios financeiros comuns, ponderada por coeficientes. Os coeficientes foram estimados através da identificação de um conjunto de empresas que tinha declarado falência e de uma amostra de empresas que sobreviveram, relacionadas industrialmente e de tamanho aproximado (activos).

No modelo *Z-Score* considera-se:

- X_1: *Working Capital*/Total de Activos
- X_2: Lucro conservado / Total de Activos
- X_3: EBIT / Total de Activos
- X_4: Mercado / Dívida
- X_5: Vendas / Total de Activos

a partir das quais se estabelece a função *"Score"* para cada devedor:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 0.999X_5 \tag{1.1}$$

e para a qual se determinam as zonas de decisão:

- $Z > 2.99$ -Zona "Segura"
- $1.8 < Z < 2.99$ - Zona "Intermédia"
- $Z < 1.80$ -Zona de "Stress"

Os modelos de *Credit Scoring* dividem-se ainda em duas categorias: os *Modelos de Aprovação de Crédito* (*Credit Scoring* propriamente dito) e os *Modelos de Classificação Comportamental*, estes últimos conhecidos como *behavioural scoring*, conforme refere [Caouette et al., 1998].

Behavioural Scoring

Segundo [Araújo, 2006], o *behavioural scoring* é uma ferramenta utilizada para previsão de eventos associados ao risco de crédito, como o incumprimento e os pagamentos em dia, entre outras características. O *behavioural scoring* tem em consideração os aspectos comportamentais e as actividades dos clientes da instituição, ou seja, auxiliam a gestão dos créditos dos clientes que já possuem créditos na instituição.

Nos modelos de *behavioural scoring* a instituição financeira analisa o comportamento dos clientes em operações anteriores. Estes modelos visam estimar a probabilidade de incumprimento de um cliente que já possui um produto ou um crédito com a mesma, enquanto que nos modelos de aprovação de crédito a instituição financeira não tem conhecimento do comportamento do cliente. O número de atrasos no último ano e o volume de transacções do cliente contribui para a análise do risco de crédito relacionado com o comportamento do cliente.

Os modelos de aprovação de crédito são orientados essencialmente para a concessão e volume de crédito, enquanto que os modelos de classificação comportamental visam principalmente a gestão dos limites de crédito, as cobranças preventivas e outras estratégias.

As técnicas de *Credit Scoring* são conhecidas e utilizadas em praticamente todas as instituições financeiras. O uso das tecnologias computacionais avançadas permite, ao usuário comum, o acesso aos poderosos sistemas de simulação, capazes de modelar a exposição ao crédito.

O desenvolvimento informático permitiu o uso de outras técnicas na construção de sistemas de *Credit Scoring*. Nos anos 80 foram introduzidas as técnicas de regressão logística e regressão linear múltipla e que são, actualmente, as duas principais técnicas utilizadas na construção dos modelos. Mais recentemente, as técnicas de inteligência artificial, como as redes neuronais artificiais, foram também implementadas com sucesso.

Vantagens dos modelos de *Credit Scoring*

[Chaia, 2003] resume as principais vantagens dos modelos de *Credit Scoring*:

- *Consistência*: são modelos bem elaborados, que utilizam a experiência da instituição, e ajudam na gestão dos créditos de clientes já existentes e de novos solicitantes;

- *Facilidade*: os modelos de *Credit Scoring* visam a simplicidade e a fácil interpretação, com instalação relativamente fácil;

- *Melhor organização da informação de crédito*: a sistematização e organização das informações contribuem para a melhoria do processo de concessão de crédito;

- *Redução da metodologia subjectiva*: o uso do método quantitativo com regras claras e bem definidas contribui para a diminuição da subjectividade da avaliação do risco de crédito;

- *Maior eficiência do processo*: o uso de modelos de *Credit Scoring* na concessão de crédito direcciona os esforços dos analistas, trazendo redução de tempo e maior eficiência a este processo.

Desvantagens e limitações dos modelos de *Credit Scoring*

Para [Silva, 2000] a principal limitação dos modelos de *Credit Scoring* é o aspecto temporal da amostra:

"O tempo (a época) é uma das principais limitações apresentadas pelos modelos desenvolvidos a partir do uso de análise discriminante. Com o decorrer do tempo, tanto as variáveis quanto seus pesos relativos sofrem alterações. As variáveis que, segundo a análise discriminante, são as que melhor classificam sob determinada conjuntura económica, podem não o ser em outra situação."

[Caouette et al., 1998] também referem o aspecto temporal como a principal limitação:

"Um modelo de *Credit Scoring* pode degradar-se com tempo se a população em que ele é aplicado diverge da população original que foi usada para construir o modelo".

[Chaia, 2003] também fala das principais desvantagens dos modelos de *Credit Scoring*:

- *Custo de desenvolvimento*: o desenvolvimento de sistemas de *Credit Scoring* acarreta custos ao nível da sua construção e manutenção.

- *Excesso de confiança nos modelos*: algumas estatísticas podem estimar por valores superiores a eficácia dos modelos, provocando um excesso de confiança nos mesmos por parte dos utilizadores, pois os menos experientes consideram-nos perfeitos e não põem em causa o seu resultado;

- *Falta de dados adequados*: a necessidade de dados não facultados pode originar problemas na utilização dos modelos e gerar resultados diferentes dos esperados. É necessário analisar a qualidade e a fidedignidade das informações disponibilizadas.

- *Interpretação equivocada das classificações*: o uso inadequado do sistema, devido à falta de treino e falta de formação sobre a sua utilização, pode provocar problemas sérios à instituição.

1.1.6 Modelos de *Credit Rating*

Segundo a Companhia Portuguesa de Rating, o *rating* é uma opinião sobre a capacidade e vontade de uma entidade vir a cumprir, de forma atempada e na íntegra, determinadas responsabilidades.

Os modelos de *Credit Rating* consistem numa classificação, publicada por agências de crédito, baseada numa análise detalhada de um histórico financeiro, referente essencialmente à própria capacidade de um cliente cumprir com as obrigações de dívida. A classificação mais alta é geralmente designada por AAA e a menor por D, como se verá adiante.

Conforme [Araújo, 2006], os modelos de *Credit Rating* baseiam-se na probabilidade de perda para classificar os empréstimos em categorias de risco de crédito, atribuindo-lhes valores e agrupando os créditos de acordo com a pontuação obtida. A atribuição de pontuação aos clientes fornece uma visão geral da capacidade financeira e do nível de risco do cliente.

A avaliação dos factores qualitativos e quantitativos relativos a um cliente tem como resultado o *rating* desse cliente. Segundo [Securato, 2002], para obter o *rating* de um cliente devem ser analisadas as condições do proponente (a indústria em que actua, a sua posição na indústria, demonstrações financeiras, restrições cadastrais, etc), a operação de crédito (prazo, garantia, liquidez da garantia) e o cenário macroeconómico no geral. O agrupamento em classes e a ponderação desses factores constitui o *rating* do solicitante de um empréstimo ou do emissor de um título.

Actualmente as instituições financeiras desenvolvem internamente os próprios modelos de *Credit Rating* ou utilizam os que são facultados por agências de *rating*. Estas agências são empresas especializadas na avaliação do risco de crédito. As agências de *rating* mais conhecidas são a *Moody's*, a *Standard and Poor's* e a *Fitch*.

Na generalidade, a maioria dos sistemas de *rating* considera até dez categorias de classificação de crédito, nas quais é estabelecida, individualmente, a percentagem esperada de *default* que irá permitir a mensuração da perda total esperada.

Os sistemas de *rating*, além da sua utilidade nas operações de crédito, são também necessários devido às exigências do Banco de Portugal, que enfatizam a classificação de risco como forma de classificação da carteira de crédito da instituição financeira.

Agency Credit Ratings

Classes de classificação ou *Rating*

A *Moody's*, a *Fitch* e a *Standard and Poor's* são agências que produzem classificações de crédito *(credit ratings)* para emitentes de instrumentos de dívida. As agências de crédito utilizam modelos registados para classificar o emitente e as obrigações do emitente numa das várias classes de classificação de crédito. Por exemplo, a *S&P* e a *Moody's* utilizam a seguinte classificação:

Avaliação de Obrigações de Longo Prazo		
S&P	Moody's	Significado
AAA	Aaa	Altíssima Qualidade, Risco de crédito mínimo.
AA	Aa	Alta Qualidade, Risco de crédito muito baixo.
A	A	Qualidade "Média Superior", Risco de crédito baixo.
BBB	Baa	Protecção adequada, Risco de crédito moderado.
BB	Ba	Propenso a pagar, mas com alguma incerteza.
B	B	Especulativas, Risco de crédito alto.
CCC	Caa	Vulnerabilidade para ocorrência de *Default*, Risco de crédito alto.
CC	Ca	Altamente Especulativas, Risco bastante alto, com algumas perspectivas de recuperação.
C	C	Muito Propenso a não-pagamento; Classe nominal mais baixa e são consideradas em *Default* pela Moody's.
D	-	*Default*
NR	-	*Non-Rated*

Os devedores com classificação BBB ou superior são usualmente chamados de *"investment-grade"* e os investidores com classificação BB ou inferior são designados como "*speculative-grade*". Cada categoria pode ainda ser seguida de um sinal de $+$ ou $-$ na escala da S&P ou dos números 1,2 ou 3 na *Moody's*, para mostrar a duração relativa dentro da categoria.

A avaliação de crédito é utilizada como um certificado independente da capacidade do emissor reembolsar as suas obrigações e empréstimos. A classificação é um factor muito útil na tomada de decisão de venda/compra de obrigações a um grande grupo de pequenos investidores.

As classificações de *rating* são revistas regularmente sendo, no entanto, necessário haver um equilíbrio entre a estabilidade e a precisão das classificações.

1.2 Medidas de Risco de Crédito

1.2.1 *Default Loss*

[Caouette et al., 1998] afirmam que o risco de crédito é a mais antiga forma de risco nos mercados financeiros, tão antigo quanto os próprios empréstimos, remontando a pelo menos 1800 a.C..

O risco de crédito é um dos riscos mais comuns numa instituição financeira, dado que a concessão de crédito é a sua principal actividade.

Segundo [Pereira, 2009], o risco de crédito é o risco de perda devido a uma falha de pagamento por parte de um devedor. Este risco está presente em todas as operações de crédito, seja na área comercial, industrial, mas principalmente nas instituições financeiras. Por essa razão, é de extrema importância a análise de risco de crédito, com intuito de identificar o nível de risco presente numa operação de crédito.

Definição 1.1. ***Risco de Crédito***
Define-se Risco de Crédito (ou Credit Risk) como sendo a possibilidade da contraparte falhar um pagamento. No instante em que essa situação ocorre diz-se que o devedor originou um default (incumprimento).

Para analisar as perdas em detalhe, considere-se a seguinte notação para cada devedor:

- τ - Instante em que ocorre o *default*;
- t - Instante no tempo que ocorre antes da maturidade ou data de vencimento, ou seja, $t \in [0, T]$;
- Indicador do processo de *Default* - $N(t) = \mathbb{1}_{\{\tau \leq t\}}$;
- Processo de Exposição *(Exposure at Default - EAD)*
  $E(t)$ - Montante que o credor perde, caso o devedor entre em *default* no tempo t, com probabilidade de recuperação igual a zero;

  Medir o risco de crédito passa também por, implicitamente, medir o montante do empréstimo não recuperável, em caso de *default* por parte do devedor. Isto significa que haverá necessidade de estimar a probabilidade de um devedor falhar o cumprimento de uma obrigação, bem como estimar o montante que a instituição financeira perderá, em caso de ocorrência de *default*.
- $R(t)$ - Taxa de Recuperação (ou *Recovery Rate*)
  É a percentagem do montante de crédito concedido que a instituição financeira recupera, em caso de ocorrer *default* por parte do devedor.
  Não é simples estimar a taxa de recuperação, pois os dados sobre as mesmas são, geralmente, fragmentados e dependem bastante dos processos de falência e da conjuntura económica à data da ocorrência do *default*.
- $L(t)$ - *Loss Given Default (LGD)*
  Representa a percentagem do montante de crédito concedido que a instituição financeira perde em caso de incumprimento de um devedor, no instante t.

  A *Loss Given Default*, $L(t)$, pode ser definida também pela taxa de recuperação, *Recovery Rate*, $R(t) = 1 - L(t)$, e é um factor importante para a *Default Loss*, $D(t)$, que se define adiante.

Em geral(ver [Pereira, 2009]), espera-se que as taxas de recuperação sobre a dívida sejam influenciadas pelos seguintes factores:

- *Garantia Real e Classe de Prioridade;*

- *Legislação local onde o processo de falência decorre;*
- *Grupo industrial do devedor;*
- *Classificação do Devedor antes do default.*
- *Ciclo Económico* (as taxas de recuperação são menores em alturas de recessão, ou seja, mais *defaults* e menores taxas de recuperação).

As taxas de recuperação dependem sistematicamente do Ciclo Económico, há poucas hipóteses para a diversificação da estimação de erros. É complicado dizer, com bastante segurança, quais as taxas de recuperação que se devem utilizar nos modelos de risco de crédito.

Defina-se agora o conceito de *Default Loss* (ver [Pereira, 2009]):

Definição 1.2. *Default Loss*
Para cada devedor, a Default Loss é o montante perdido pelo credor devido à ocorrência de default e é dada pela expressão seguinte:

$$D(T) := N(T) \times E(\tau_i) \times L(\tau_i) \tag{1.2}$$

onde T é a maturidade (considera-se, por exemplo, 1 ano) e τ o instante em que ocorre o default para o devedor i.

1.2.2 Probabilidade de *Default*

Segundo [Pereira, 2009], a Probabilidade de *Default* determina o processo $N(t)$ referido em (1.2), logo, pode dizer-se que a probabilidade de *default* é uma das componentes da *Default Loss*.

Seja $PD_i(t, T)$ a probabilidade de *default* de um devedor i, deste o instante t até T.

A probabilidade de *default* pode ser representada por uma árvore binomial.

Considerando a árvore binomial a um período, por exemplo, o devedor pode estar em dois estados possíveis: Estado *Default* (D_1) com probabilidade p_1, ou estado de sobrevivência (S_1) com probabilidade $1 - p_1$. Para o caso a n-períodos procede-se de forma análoga. Neste caso, pode observar-se na Figura 1.1 que existem n caminhos distintos que levam ao *default*, mas apenas um que leva ao *non-default*, ou sobrevivência.

Denote-se $PS(t, T)$ a probabilidade de sobrevivência desde o instante t até T.

- A probabilidade cumulativa de sobrevivência desde $t = 0$ até ao período n é:

$$PS(0, n) = P[\text{sobreviver de 0 a } n] = (1 - p_1)(1 - p_2) \dots (1 - p_n).$$

- A probabilidade cumulativa de *default* até ao período n é a probabilidade de entrar em incumprimento em qualquer instante no tempo antes do período n, e é dado por:
$$PD(0,n) = 1 - PS(0,n).$$

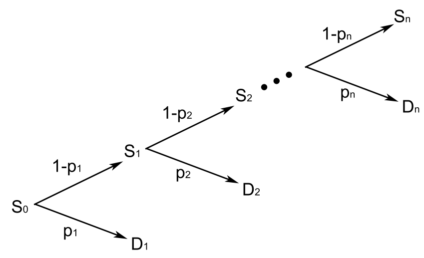

Figura 1.1: Árvore Multi-Períodos

Perdas Esperadas

Seja $PD_i(0,T)$ a probabilidade de *default* do devedor i.
Temos, conforme refere [Pereira, 2009], que:

$$\mathbb{E}[N_i(T)] = PD_i(0,T).$$

Definição 1.3. ***Perda Esperada Individual***
Sejam E_i e L_i conhecidos e constantes. A perda esperada de um devedor durante o tempo T é:

$$\mathbb{E}[D_i(T)] := PD_i(0,T) \times E_i \times L_i. \tag{1.3}$$

Note-se que embora a perda esperada individual possa ser pequena, as perdas acumuladas, quando consideramos uma carteira de vários devedores, pode ser muitíssimo grande.

1.2.3 Matrizes de Transição e Migração de Crédito

Transição a 1-Período

Segundo [Pereira, 2009], em modelos quantitativos de gestão de risco de crédito, a utilização das classificações referidas anteriormente na Secção 1.1.6, está sujeita a alterações. Mais precisamente, é necessário fazer corresponder as classificações alfabéticas em probabilidades de migração de crédito. Isto é, é necessário atribuir valores à probabilidade de um emitente se mover nas classes de risco ou mesmo entrar em *default*. As agências de crédito também publicam este tipo de informação em matrizes de transição.

A Moody's publicou a matriz ilustrada na Tabela 1.1

Initial Rating	**Rating at year-end (%)**							
	Aaa	**Aa**	**A**	**Baa**	**Ba**	**B**	**Caa**	**Default**
Aaa	93.40	5.94	0.64	.000	0.02	0.00	0.00	0.00
Aa	1.61	90.55	7.46	0.26	0.09	0.01	0.00	0.02
A	0.07	2.28	92.44	4.63	0.45	0.12	0.01	0.00
Baa	0.05	0.26	5.51	88.48	4.76	0.71	0.08	0.15
Ba	0.02	0.05	0.42	5.16	86.91	5.91	0.24	1.29
B	0.00	0.04	0.13	0.54	6.35	84.22	1.91	6.81
Caa	0.00	0.00	0.00	0.62	2.05	4.08	69.20	24.06

J.P.Morgan,(1997, tabela 6.1)

Tabela 1.1: Matriz de Transição

Cada entrada da matriz representa a probabilidade de migração da classe-linha para a classe-coluna.

Note-se que a soma das probabilidades de cada linha soma 1, pois uma empresa que esteja com classificação *Baa*, por exemplo, estará certamente em alguma das classes possíveis no próximo ano.

Os elevados valores na diagonal da matriz reflectem a "estabilidade da classificação" determinada pelas agências de classificação.

Caso se acrescente uma linha à matriz de transição, destinada ao estado de *default*, esta linha só teria zeros, exceptuando na última coluna que teria o valor 100%. Assume-se, portanto, que o estado *"default"* é um estado absorvente.

Transição a N-Períodos

Considerando a matriz de transição a 1-período, pode ir-se mais além e produzir uma matriz de transição a n-períodos. Para tal, (ver [Pereira, 2009]) consideram-se duas suposições:

- *Tempo de invariância.* Assume-se que a matriz de transição a 1-período é constante, ou seja, não depende do tempo de calendário.

- *Propriedade de Markov.* Probabilidades de migração de classificação só dependem da avaliação actual. A classificação histórica (*"upgrade"/ "downgrade"*) é irrelevante para a determinar estados futuros.

Estas duas suposições estão sujeitas a alguma crítica. Primeiro, existe alguma evidência que o "downgrading" é mais provável em situações de recessão do que em fases de "boom" do ciclo económico. Em segundo lugar, existe evidência de que o momento da classificação é importante, ou seja, um "downgrade" recente torna mais provável um novo "downgrade" para uma mesma empresa, por exemplo, do que para uma que se mantenha na mesma escala de classificação há algum tempo. No entanto, estas duas suposições são consideradas desde que permitam a simplificação dos modelos de risco de crédito.

Essas suposições permitem que a análise da evolução temporal da classificação, em termos de risco de crédito, possa ser tratada recorrendo à sobejamente conhecida Teoria das Cadeias de Markov.

Para se obter a probabilidade de transitar do estado i para o estado j, em dois períodos, há que considerar todos os caminhos possíveis que conduzem de i a j. A probabilidade é a soma de todas as possibilidades possíveis, $p_{ij}^{(2)} = \sum_{k=1}^{K} p_{ik}p_{kj}$, onde $K = 8$ para a matriz de transição da *Moody's* na tabela anterior. Então a matriz de transição a 2-períodos será $P^{(2)} = P \times P$.

Em geral, a matriz de transição a n-Periodos é dada por:

$$P^{(n)} = P^n \tag{1.4}$$

As classificações referidas anteriormente estão normalmente disponíveis apenas para grandes empresas. No entanto, um banco tem pequenas e médias empresas na sua carteira. Para a avaliação de pequenas empresas, foram desenvolvidos vários métodos estatísticos. Os modelos de classificação de crédito podem utilizar a seguinte informação para prever a probabilidade de *default*:

- Dados com balanço sobre endividamento do devedor;
- Dados com balanço sobre a capacidade de pagamento;
- Grau de risco do negócio. Dados sobre a volatilidade;
- A capitalização de mercado da empresa;
- Dados macro-económicos.

1.3 Modelos para Carteiras

A estimação da probabilidade de *default* para um cliente individual é uma análise importante a efectuar por parte da instituição financeira, mas será mais relevante do ponto de vista da concessão, ou não, do crédito solicitado. No entanto, a importância da probabilidade de *default* vai muito além da análise de concessão de crédito.

As implicações futuras de um eventual *default* reflectir-se-á no desempenho e análise da carteira de clientes da instituição. Assim sendo, as Instituições financeiras necessitam de instrumentos que analisem a carteira de clientes no seu todo. Por exemplo, é através da análise conjunta dos clientes que serão estimadas as reservas de capital necessárias para fazer face ao risco de crédito da carteira.

Na situação que será tratada nesta secção, a variável aleatória de interesse será a *Default Loss* de uma carteira.

1.3.1 *Default Loss* de uma Carteira

Definição 1.4. ***Portfolio Default Loss***
A Default Loss, numa carteira com J devedores, é definida por (ver [Pereira, 2009]):

$$D(T) := \sum_{j=1}^{J} D_j(T) = \sum_{j=1}^{J} N_j(T) \times E_i(\tau_j) \times L_j(\tau_j), \tag{1.5}$$

com $D_j(T)$, $N_j(T)$, $E_i(\tau_j)$ e $L_j(\tau_j)$ definidos como na Subsecção 1.2.1.

Distribuição da *Default Loss* de uma carteira

A distribuição do Total das perdas de crédito da carteira pode ser representada por $F(x) := P[D(T) \leq x]$. Na generalidade das situações, a assumpção de independência entre as perdas individuais reflecte-se numa sub-avaliação do risco de crédito da carteira (ver [Pereira, 2009]).

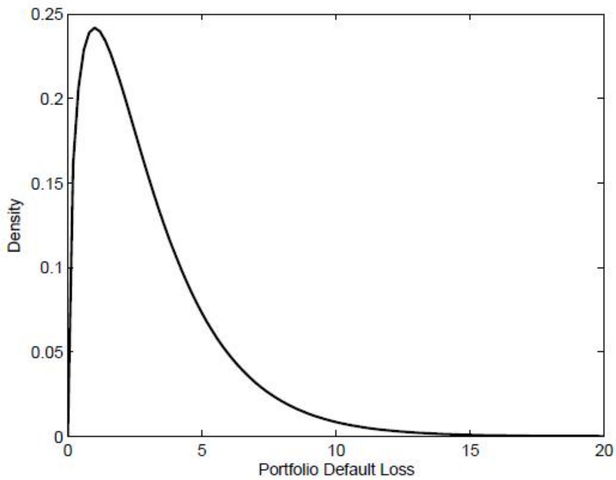

Figura 1.2: Função Densidade Típica da *Default Loss* de Uma Carteira

O gráfico da função densidade da *Default Loss* de uma carteira de empréstimos usuais tem, regra geral, as seguintes características:

- *Não simétrico*: As perdas são truncadas em zero, $D \geq 0$, (o melhor caso é quando não há perdas) e, no entanto, a perda pode ser extremamente grande.

- *Muito inclinada positivamente*: A maior massa está na área de pequenas perdas, mas existe uma cauda longa para a direita.

- *Cauda direita pesada*: as grandes perdas são ainda prováveis. Os quantis do *Value-at-Risk*[1] (VaR) usual são assim associadas a perdas muito grandes.

Intensidade de *Default*

Segundo [Pereira, 2009], uma classe importante de modelos de crédito assume que a ocorrência de *default* é um acontecimento aleatório.

As hipóteses usuais são que o número de *defaults* (N), numa carteira de crédito, durante um certo intervalo de tempo, segue uma distribuição de Poisson.

Seja λ a probabilidade de *default* por unidade de tempo (geralmente 1 ano). Esta quantidade é designada por taxa de incumprimento (*"default rate"*, *"default intensity"* ou *"default hazard rate"*).

Da definição de distribuição de Poisson temos que a probabilidade de ocorrerem k eventos em Δt é:

$$P[N = k] = \frac{e^{-\lambda \Delta t}(\lambda \Delta t)^k}{k!} \text{ para } k \in \mathbb{N}_0.$$

Assim, a probabilidade de sobrevivência é dada pela probabilidade de não ocorrer *default*, i.e.:

$$PS(t, t + \Delta t) = P[N = 0] = e^{-\lambda \Delta t}.$$

Em geral, podemos ter diferentes intensidades de *default* para os diferentes períodos. Se os modelos de crédito são formulados em tempo contínuo será vantajoso definir uma probabilidade *instantânea* de *default*.

Seja $\lambda(t)$, para $t \geq 0$, a probabilidade de *default* por unidade de tempo (1 ano, geralmente) para o instante t. Considera-se o limite em tempo contínuo fazendo $\Delta t \to 0$. [Pereira, 2009] refere que a probabilidade de sobrevivência, desde t até T, é dada pelo Processo de Poisson:

$$PS(t, T) = exp\{-\int_t^T \lambda(s)ds\}.$$

Perdas Esperadas de uma carteira

A perda esperada de uma carteira pode ser definida, segundo [Pereira, 2009], como:

Definição 1.5. ***Perda Esperada de uma Carteira***

$$\mathbb{E}[D(T)] := \sum_{i=1}^{I} \mathbb{E}[D_i(T)]. \tag{1.6}$$

Quando se concede um empréstimo, os ganhos esperados devem cobrir a perda esperada do respectivo empréstimo. O princípio é idêntico ao princípio que rege a Actividade Seguradora.

[1] Ver Glossário

Perdas Inesperadas de uma Carteira

Para [Pereira, 2009], as perdas inesperadas são medidas em relação ao quantil VaR e são como uma medida de um cenário imprevisível (cenário de stress).

Seja D_k o quantil k do VaR da carteira, isto é, a perda máxima com nível de confiança k.

Definição 1.6. ***Perda Inesperada de uma Carteira***
A perda inesperada de uma carteira (Unexpected Portfolio Loss - UEL) é a diferença entre o quantil do VaR e a perda esperada da mesma carteira:

$$UEL := D_k - E[D(T)] \tag{1.7}$$

Esta não é, no entanto, a perda máxima possível. A perda máxima possível ocorre quando o total da carteira se perde e não há possibilidade de recuperação. Ao contrário das perdas esperadas, as inesperadas não são aditivas em relação às perdas individuais. A perda inesperada é utilizada com frequência para determinar a reserva de capital necessária para cobrir o risco de crédito da carteira.

Podemos concluir destes resultados que:

- Perdas inferiores à perda esperada são suportadas por linhas individuais de crédito.
- Perdas adicionais superiores à perda esperada, mas inferiores à perda inesperada, são suportadas pelas reservas de capital.
- Finalmente, as perdas superiores à perda inesperada podem resultar em falência da instituição financeira. No entanto, este tipo de acontecimento pode ser controlado, definindo o nível original do VaR, ou seja alterar o valor da pior perda esperada de modo a que sejam tomadas medidas para que não existam perdas superiores às perdas esperadas e inesperadas.

1.3.2 Exemplos de Modelos para Carteiras

Segundo [Pereira, 2009], o risco de um devedor individual é o ponto de partida para determinar a taxa de juro a cobrar por um empréstimo ou obrigação. No entanto, a prioridade dos bancos é o risco de crédito da sua carteira de crédito global. A estabilidade do banco, depende não de cada devedor, mas sim do conjunto que forma a sua carteira de crédito, isto é, depende do desempenho global da sua carteira de crédito.

Um dos factores mais importantes do risco de crédito numa carteira é o risco de concentração (*concentration risk*). Se a carteira está concentrada numa área geográfica restrita ou numa só área industrial, a instituição financeira fica muito mais exposta a choques económicos.

Os factores-chave nos modelos de risco crédito para carteiras são as correlações existentes entre as contrapartes. No entanto, a estimação das correlações não é fácil devido à escassez de dados. Para que seja possível aplicar os modelos, é necessário simplificar alguns pressupostos.

Os retornos de crédito (retornos de uma carteira de empréstimos ou obrigações) não são normalmente distribuídos, ao contrário dos retornos de mercado, o que torna ainda mais complicado a obtenção das correlações. A assimetria da distribuição é causada pelas perdas de crédito. Assim, a média e a variância não serão estatísticas suficientes para compreender a distribuição dos retornos de crédito. A Figura 1.3 ilustra uma típica distribuição de Retornos de Crédito e compara-a com uma distribuição típica de Retornos de Mercado.

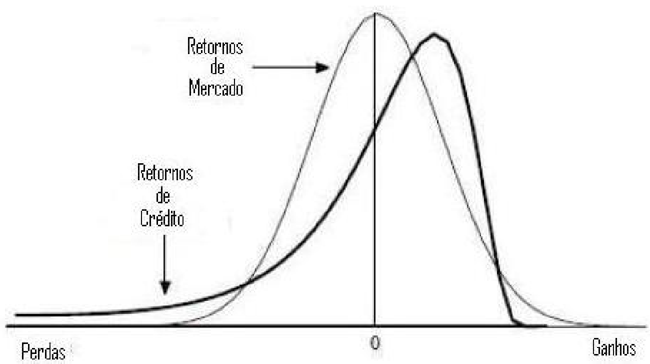

Figura 1.3: Distribuição Tipíca de Retornos de Crédito e Retornos de Mercado

Nesta sub-secção serão apresentados alguns modelos de risco de crédito para carteiras de clientes. Todos os modelos se concentram na estimativa das medidas de risco de crédito para carteiras, embora utilizem diferentes técnicas.

Segundo [Neto, 2002], as recentes abordagens sobre gestão de risco de crédito estão interligadas aos modelos para carteiras. Os modelos para carteiras têm a sua origem na teoria da diversificação de Harry Markowitz (1952). A ideia inicial de Markowitz era que o risco de um título podia ser caracterizado pela sua variância e que o retorno de uma carteira ou grupo de títulos dependia da variância dos retornos individuais dos títulos e ainda da covariância entre os mesmos.

O estudo apresentado por Harry Markowitz baseia-se num modelo de diversificação que propõe a redução do risco até perto do risco sistemático. O risco sistemático é o risco inerente ao sistema, sendo influenciado por factores/decisões macroeconómicas.

A diversificação faz com que o risco total da carteira seja menor que a soma dos riscos individuais de cada activo. Assim, parte do risco total do activo é eliminado quando este é incluído na carteira, sendo que o risco que resta representa a contribuição de risco do activo na carteira total. A diversificação não deve, no entanto, ser supérflua ou aleatória. Uma diversificação "máxima" pode reduzir desnecessariamente o retorno da carteira e os investidores preferem carteiras com retorno maior e risco maior ou retorno menor e risco menor. Perante um mercado eficiente, o investidor deveria preocupar-se com a relação título/carteira.

A aplicação da teoria de carteiras à avaliação de risco de crédito resulta no risco da carteira como um todo. Assim sendo, os modelos de risco de crédito de carteiras diferenciam-se dos restantes modelos de risco de crédito por terem em conta o risco dos activos, mas também o risco da carteira em geral. Por outras palavras, focalizam a análise de crédito individual e a análise da carteira de crédito na globalidade. Em suma, o principal objectivo de um modelo de risco de crédito de carteiras é obter a distribuição de perdas de uma carteira resultante do incumprimento ou da desvalorização da mesma num horizonte temporal determinado. Os modelos de risco de crédito de carteiras mais divulgados no mercado são: *CreditMetrics*, *CreditRisk+*, *CreditPortfolioView* e *KMV*.

No que se segue, será feito uma descrição sumária destes modelos.

CreditMetrics

Em Abril de 1997, o Banco J.P. Morgan anunciou o novo modelo *CreditMetrics* (ver [Gupton et al., 1997]), um modelo de avaliação de risco, específico para risco de crédito. Para [Securato, 2002], este modelo tinha como objectivo a administração do risco total das carteiras de crédito através da metodologia VaR - Valor em Risco.

[Pereira, 2009] refere que este método consiste num método de migração de crédito que estima as alterações futuras do valor de uma carteira de empréstimos e obrigações, que se devem a mudanças de classificação dos devedores de crédito. O método baseia-se nas categorias de classificação usuais e a respectiva matriz de transição (veja-se a matriz de transição 1.1 na secção 1.2.3).

Segundo [Araújo, 2006], o VaR pode ser definido como uma medida da maior ou pior perda esperada numa carteira de activos, dentro de um período de tempo e intervalo de confiança estipulado. O VaR é a estimativa de perda máxima que uma carteira pode apresentar durante um determinado período de tempo. Baseia-se no comportamento dos activos que a constituem e é calculada em função das volatilidades e correlações anteriores.

O VaR, como medida de risco, começou a ser utilizado na área do risco de crédito após a criação do *CreditMetrics*, até então era apenas utilizado para mensurar o risco de mercado.

Dentro da análise e avaliação do risco de crédito, esta metodologia trata as mudanças de qualidade de crédito, desenvolvendo três componentes básicas:

1. Definição do valor exposto ao risco de crédito;

2. Mensuração das volatilidades do valor devido à mudança da qualidade do crédito;

3. Medição das correlações entre os activos

O *CreditMetrics* procura identificar, além da probabilidade de *default*, o *Value-at-Risk* (VaR) da carteira de empréstimos num horizonte que inclui valorizações e desvalorizações da qualidade de crédito e as possíveis migrações de classificação (*rating*). O objectivo do *CreditMetrics* é estimar a distribuição do valor de qualquer carteira de activos sujeitos a alterações de qualidade de crédito.

O *CreditMetrics* oferece, assim, uma abordagem mais ponderada aos limites de crédito.

Segundo [Pereira, 2009], para simplificar foram definidos os pressupostos:

- A única fonte de incerteza é a migração de crédito, ou seja, os movimentos na escala de classificação.
- Dentro da mesma classe todos os emitentes são homogéneos, isto é, todos eles têm as mesmas probabilidades de transição e de *default*.
- As taxas de juro são deterministas (como se não houvesse incerteza). Assim, o valor futuro dos títulos é calculado utilizando as taxas de hoje.
- O valor das acções da empresa *(Equity value)* é um bom substituto para o valor dos activos da empresa. Assim, as correlações entre os valores dos activos podem ser aproximadas por correlações entre os preços das acções. Na verdade, as acções (*equity*) são mais voláteis que os activos devido à alavancagem.

O método da CreditMetrics utiliza dois tipos principais de construção:

- Risco de crédito para um único instrumento.
- Risco de crédito para toda a carteira, que representa as correlações entre os eventos de crédito.

Para estimar a alteração do valor de uma única obrigação durante o espaço de 1 ano, é necessário calcular o preço da obrigação para cada migração de crédito possível. Para calcular as medidas de risco e o risco de crédito de uma determinada obrigação segue-se os passos seguintes:

1. Estimar/obter a matriz de transição. Pode ver-se, na Tabela A.1 em anexo, um exemplo de uma matriz de transição.
2. Estimar a taxa de recuperação para as antigas características da obrigação.
3. Utilizar as taxas de juro para cada categoria de classificação para avaliar a obrigação em cada estado. V é o valor da obrigação.
4. Cálculo de medidas de risco de crédito.

O valor esperado da obrigação é o valor médio de todos os estados:

$$\mu_V \equiv \sum_s p_s V_s,$$

onde p_s é a probabilidade de pertencer ao estado s.

Deve ter-se em conta que o CreditMetrics considera as perdas originadas por *default* e também por *downgrades*.

Uma medida alternativa de risco também calculado pelo CreditMetrics é o desvio padrão do valor da obrigação, obtido a partir de:

$$\sigma_V^2 \equiv \sum_s p_s(V_s - \mu_V)^2 = \sum_s p_s(\Delta V_S)^2.$$

CreditRisk+

O *CreditRisk+* foi desenvolvido pela Credit Suisse Financial Products *(CSFP)*, através da aplicação de ideias já utilizadas na área dos seguros para modelar o risco de crédito.

O *CreditRisk+* é um modelo estatístico para risco de *default* que considera a taxa de incumprimento como uma variável aleatória contínua. Neste modelo não são feitas suposições sobre as causas de incumprimento, consideram-se apenas cenários de incumprimento, com probabilidade p, e de cumprimento com probabilidade $1 - p$. Temos então que:

- Para um empréstimo ou título, a probabilidade de *default* num dado período, um ano, por exemplo, é igual em qualquer outro ano;
- Para um grande número de empréstimos ou títulos, a probabilidade de *default* de um título ou empréstimo em particular é pequena, e o número de *defaults* que ocorrem num dado período é independente do número de *defaults* que ocorrem em qualquer outro período.

Partindo destes pressupostos, a distribuição da probabilidade do número de *defaults*, N, numa carteira e num dado período de tempo, pode ser aproximada por um Processo de Poisson Homogéneo:

$$P[\text{N=n}] = \frac{\mu^n e^{-\mu}}{n!}, \text{ para } n \in \mathbb{N}_0$$

onde μ é o valor médio do número de *defaults* por período de tempo (normalmente 1 ano).

Segundo [Pereira, 2009], o *CreditRisk+* utiliza este modelo básico de estatística, para derivar expressões de forma fechada para a distribuição da perda de uma carteira.

O *CreditRisk+* apenas incide sobre a ocorrência de *default*, o que tem vantagens e desvantagens. A principal vantagem é que é fácil de implementar e é computacionalmente atractivo. A principal desvantagem é que ignora o risco de migração (desclassificação - *downgrading*).

CreditPortfolioView

O modelo CreditMetrics considera, como se viu anteriormente, uma matriz de transição constante. A matriz é incondicional, no sentido em que é uma média histórica baseada em muitos anos de dados, relativos a vários ciclos económicos. No entanto, sabe-se que a ocorrência de *default* e as probabilidades de migração de crédito estão ligadas ao estado da economia, com os ciclos de crédito a seguir-se perto dos ciclos económicos. Em períodos de recessão, as probabilidades de *default* e *downgrade* aumentam; em períodos de expansão verifica-se o oposto.

Wilson (1997a, b) propõe um aperfeiçoamento do método de migração de crédito denominada *CreditPortfolioView*, que está agora comercialmente implementado pela McKinsey.

Neste método, as probabilidades de transição variam ao longo do tempo. Mais precisamente, elas são condicionais ao estado da economia.

O modelo *CreditPortfolioView* utiliza um índice para descrever o estado da economia. O índice é função de variáveis macro-económicas, tais como:

- Taxa de desemprego;
- Taxa de crescimento do PIB;
- Taxas de juro;
- Taxas de câmbio;
- Gastos do governo;
- Taxa de poupança agregada.

As probabilidades condicionadas de transição são estimadas como funções deste macro-índice.

A limitação desta abordagem é que, sendo condicional, exige uma maior quantidade de *defaults* e migração de dados para calibrar o modelo.

O modelo *CreditPortfolioView* é considerado um modelo multi-factor utilizado em simulação de distribuições condicionais de probabilidade de migração de *rating*, ou de incumprimento, para diferentes conjuntos de empresas ou países, que estão condicionadas a factores macroeconómicos, como a taxa de desemprego, taxas de juro e câmbio, etc. Assim, a metodologia deste modelo interliga os factores macroeconómicos à probabilidade de *default* e migração da qualidade de crédito.

Neste modelo, os devedores são agrupados em sectores definidos por classes económicas, países e *ratings*.

Esta abordagem constrói um modelo econométrico que procura explicar e determinar a proabilidade de incumprimento para os diferentes sectores, baseando-se em variáveis macroeconómicas históricas e em séries temporais de taxas médias de incumprimento.

KMV

O modelo *KMV* foi criado pela *KMV Corporation* para estimar a probabilidade de *default* de empresas e baseia-se na abordagem de Merton de 1974 (pode ver-se em [Arvantis e Gregory, 2001]), na qual uma empresa é considerada em insolvência quando o valor dos seus activos é inferior ao valor de seus passivos, sendo que, a probabilidade de *default* é dada pela diferença entre o valor de mercado do activo e o valor oficial dos passivos, ver [Caouette et al., 1998].

Para aplicação deste modelo é necessário começar por estimar o valor de mercado do activo da empresa, pois este não é observável. Considera-se o capital próprio da empresa como uma opção de compra sobre os seus activos e estabelece-se uma relação entre os dados de mercado, observáveis em relação às acções da empresa, com os valores não observáveis dos activos.

O objectivo principal do KMV é apreçar o crédito, considerando a empresa como uma opção de compra.

Todos os modelos de crédito para carteiras discutidos nesta subsecção parecem ser diferentes, no entanto, vários estudos comparativos têm mostrado que, quando estes modelos são executados utilizando parâmetros consistentes, os resultados obtidos são bastante semelhantes.

Nesta secção, foi realizado um resumo teórico dos principais modelos de gestão de risco de crédito que são utilizados actualmente em instituições financeiras com o intuito de fazer um enquadramento geral dos modelos e técnicas utilizadas para modelação de risco de crédito. Ressalte-se que, não tendo como objectivo esgotar o assunto, foram abordados apenas alguns aspectos mais relevantes desses modelos.

Capítulo 2

Risco e *Spread*

Spread é por definição a diferença entre o preço de compra e o preço de venda de uma mesma acção, título ou garantia. O *spread* bancário, é a diferença entre a taxa de juro que as instituições financeiras pagam na aquisição do dinheiro e a que cobram dos clientes. É também conhecido como "taxa de risco".

O *spread* deverá representar, para além do lucro das instituições financeiras, uma medida do risco de crédito que um determinado cliente representa para a instituição financeira permitindo a compensação do risco de crédito assumido pela instituição.

Tendo por base esta ideia pretende-se, neste capítulo, obter expressões simples que definam o *spread* a aplicar a cada cliente, em função da probabilidade de incumprimento (*default*) e da taxa de recuperação *(recovery rate)*.

Todavia, determinar o *spread* para cada cliente de uma dada carteira de crédito não é simples do ponto de vista prático. Admite-se que, para essa carteira, há dados relativos a factores de risco bem como dados sobre as probabilidades de *default* e das taxas de recuperação.

O objectivo nas secções seguintes é estabelecer, em casos particulares, uma fórmula que permita determinar o *spread* em função da taxa de recuperação e da probabilidade de *default*. Observamos que, de acordo com a prática habitual (ver [McNeil et al., 2005]), existem duas metodologias de apreçamento de risco de crédito: A metodologia actuarial e a metodologia financeira ou de apreçamento através de uma medida neutra face ao risco. Estas metodologias são diferenciadas, respectivamente, pelo uso da medida de probabilidade natural e da medida martingala equivalente.

Com base no modelo discreto a um período desenvolve-se, através das metodologias actuarial e financeira, uma expressão que define o *spread* mínimo a aplicar a um dado cliente em função do risco que este representa.

Posteriormente desenvolve-se a mesma expressão com base no modelo contínuo, através da metodologia actuarial.

A expressão obtida é também estabelecida, em estudos sobre risco de crédito, por diversos autores, nomeadamente, [Arvantis e Gregory, 2001] e [Pereira, 2009]. Ambos os desenvolvimentos serão apresentados neste capítulo, com o intuito de reforçar a

importância dos resultados obtidos e ainda de forma a possibilitar algumas analogias.
Alguns dos conceitos mais importantes deste capítulo estão definidos no Glossário.

2.1 Variáveis e Parâmetros

Em qualquer abordagem sobre modelação de risco de crédito, é necessário considerar algumas definições, variáveis e parâmetros básicos.

Considere-se que:
Uma obrigação de cupão zero (*Zero coupon bond*) é um contrato de dívida que, na maturidade T, paga uma quantia (nominal) no valor de uma unidade monetária, e não paga cupões.

Considerem-se:

- $(X_t)_{t \geq 0}$- Variável que descreve a evolução do preço de um activo e que se considera ser um Processo de Markov[1];

- τ - Momento da ocorrência de *default* que deve ser um tempo de paragem. Associada a esta variável ter-se-á a probabilidade de *default*, que deverá ser um parâmetro do modelo;

- $R(T)$ - Taxa de Recuperação e que também se considera ser um processo de Markov.

  Note-se que a informação disponível sobre a taxa de recuperação nem sempre é fiável, o que sugere que este processo deva ser considerado como um parâmetro constante (os valores de R registados por uma instituição financeira referem-se apenas ao valor recuperado pela mesma, no entanto, as empresas encarregues de recuperar o património podem recuperar $k > R$ ou $k < R$ mas devolvem apenas à instituição financeira a parte acordada com a mesma).

- $B(t, T)$ - Preço, no instante t, de uma obrigação de cupão zero livre de incumprimento (*default-free zero coupon bond*) com vencimento em T;

- $\tilde{B}(t, T)$ - Preço, no instante t, de uma obrigação de cupão zero com possibilidade de incumprimento (*defaultable zero coupon bond*) com vencimento em T;

- $s(t, T)$ - *Spread* de crédito no instante t, e T a data de vencimento do contrato de crédito. Considera-se que o *spread* também deverá ser um processo de Markov.

[1]Genericamente, um processo é de Markov se os estados anteriores são irrelevantes para a predição dos estados seguintes, desde que o estado actual seja conhecido. As definições formais podem ser lidas em [Oksendal, 2007].

- Adicionalmente deve considerar-se também a existência de alguma estrutura temporal (*term structure*), isto é, considerar-se a existência de relações entre as taxas de juro, por exemplo, de duas *zero coupon bonds* com diferentes maturidades.

Seja $B(0,T)$, o preço, hoje, de uma *zero coupon bond* livre de incumprimento (*risk free*) e com maturidade T. Inerente a uma taxa de juro fixa, r, e tendo em conta a inexistência de arbitragem, pode-se representar o preço desta *zero coupon bond* como em (2.1).

$$\begin{aligned} & B(0,T) \times e^{(rT)} = 1 \\ \Leftrightarrow \quad & B(0,T) = e^{(-rT)} \end{aligned} \tag{2.1}$$

Da mesma forma, o preço, em $t > 0$, de uma *zero coupon bond* livre de risco (*risk-free*) e com maturidade T com taxa de juro fixa r, será:

$$\begin{aligned} & B(t,T) \times e^{r(T-t)} = 1 \\ \Leftrightarrow \quad & B(t,T) = e^{-r(T-t)} \end{aligned} \tag{2.2}$$

Analogamente, segundo [McNeil et al., 2005], tendo em conta que o preço de uma *defaultable zero coupon bond* tem de ser menor que o preço de *zero coupon bond* livre de risco (para compensar o respectivo risco), o preço de uma *defaultable zero coupon bond* com *spread* de crédito, $s(t,T)$, no tempo t e com vencimento em T, é tal que:

$$\tilde{B}(t,T)e^{(T-t)\times s(t,T)} = B(t,T).$$

A expressão anterior pode ser reescrita de forma a obter-se o processo que descreve o *spread*:

$$s(t,T) = -\frac{1}{T-t} ln\left(\frac{\tilde{B}(t,T)}{B(t,T)}\right) . \tag{2.3}$$

2.2 Apreçamento de Risco de Crédito

Nesta secção, discutem-se e comparam-se as metodologias actuarial e financeira, com vista ao apreçamento de títulos de crédito com risco. Apresenta-se uma relação entre o mundo real ou medida física, que modela a real probabilidade de *default*, e uma medida equivalente martingala ou medida neutra face ao risco .

2.2.1 Modelo Discreto a Um Período

A Metodologia Actuarial

Na abordagem actuarial, segundo [McNeil et al., 2005], os preços são calculados como a soma do *pay-off* esperado, sob a medida física e um prémio de risco. A dimensão do prémio de risco é muitas vezes relacionado com a noção de capital económico[1]. Em risco de crédito, a abordagem actuarial é aplicada principalmente ao apreçamento de empréstimos não negociados (*non-traded*) ou de produtos estruturados relacionados com títulos sem liquidez. Seja $(\Omega, \mathcal{F}, \mathbb{P})$, o espaço de probabilidade em que $\mathbb{P}$ é a medida natural observável.

Ilustração:

Considere-se uma obrigação ou um empréstimo de $100u.m.$, sujeitos a uma taxa de juro constante r, com probabilidade de *default* igual a p, uma taxa de recuperação R e um *spread* s.

No caso de uma obrigação livre de incumprimento (*default-free*) ou de um empréstimo, as $100u.m.$, na maturidade, converter-se-ão em $100(1+r)u.m.$. A ausência de risco de crédito implica que o *spread* e a taxa de recuperação não sejam necessários. No caso de um empréstimo ou obrigação *defaultable* (sujeitos ao risco de *default*) há que ter em consideração a probabilidade de *default*, a taxa de recuperação e o *spread*. Assim, com probabilidade $(1-p)$ (não ocorrer *default*) as $100u.m.$ converter-se-ão em $100(1+r)(1+s)u.m.$ e com probabilidade p converter-se-ão em $100R(1+r)(1+s)u.m.$.

A Tabela 2.1 ilustra o valor esperado dos *cash-flows* do empréstimo ou obrigação nos dois cenários distintos.

Time	Default-free	Defaultable
$t=0$	100	100
$t=1$	$100(1+r)$	$[100(1-p)+R100p]\,(1+r)(1+s)$

Tabela 2.1: Cash-Flows a Um Período

Nota 2.1.
Note-se que, como $R<1$, temos que

$$100(1-p)+R100p<100$$

pelo que o spread representa uma taxa de prémio de risco que relaciona o investimento de risco com o investimento sem risco.

Em termos de expectativas de mercado, deverá ter-se que:

$$\mathbb{E}[\text{ cash-flow de uma } \textit{Default-free}] = \mathbb{E}[\text{ cash-flow de uma } \textit{Defaultable}], \tag{2.4}$$

[1] Capital necessário para o banco permanecer solvente em caso de perda extrema.

ou seja,

$$100(1+r) = \Big[100(1-p) + R100p\Big](1+r)(1+s), \tag{2.5}$$

caso contrário, ninguém investiria na obrigação menos rentável.

Considere-se então o seguinte teorema:

Teorema 2.1.
No modelo discreto a um período, para o risco de crédito, o spread é função da taxa de recuperação e da probabilidade de default na medida natural de probabilidade $\mathbb{P}$ e é dado por:

$$s = \frac{(1-R)p}{1-(1-R)p} \ . \tag{2.6}$$

Demonstração. Basta resolver a Equação (2.5) em ordem a s. □

Proposição 2.1.
Para $(1-R)p << 1$ *temos um resultado clássico:*

$$s \approx (1-R)p. \tag{2.7}$$

Demonstração. Deduz-se a partir da Equação (2.6). □

Fazendo a análise, através da Metodologia Financeira, o objectivo é obter um resultado comparável.

A Metodologia Financeira

Sob a abordagem de apreçamento, [McNeil et al., 2005] refere que, com base na probabilidade neutra face ao risco, os preços são calculados conforme os valores esperados descontados sob alguma medida martingala[2] equivalente. Esta abordagem é baseada nas noções de ausência de arbitragem e de *hedging dinâmico*.

Hoje em dia, a aproximação para o apreçamento, com base na probabilidade neutra face ao risco, é *standard* para apreçar títulos *non-defaultables*. Em risco de crédito, é usado para apreçar títulos negociados, nomeadamente, *corporate bonds*, *default swaps* de crédito e títulos derivados semelhantes a estes produtos.

Uma medida martingala equivalente ou uma medida de probabilidade neutra face ao risco, é uma medida artificial $\mathbb{Q}$ equivalente à medida de probabilidade física $\mathbb{P}$, de tal forma que o processo de preços descontados do activo relevante é uma $\mathbb{Q}-$*martingala*. De acordo com um resultado de matemática financeira (o designado, *First Fundamental Theorem of Asset Pricing*), um modelo para os preços dos títulos é livre de arbitragem

[1] Genericamente, uma martingala é um processo estocástico sem tendência, que descreve um jogo equilibrado. A definição formal pode ser consultada em [Oksendal, 2007].

se e só se admite uma medida de martingala equivalente $\mathbb{Q}$. No modelo discreto, $\mathbb{Q}$ é dada simplesmente pela probabilidade de *default* artificial q.

Considere-se novamente um empréstimo ou uma obrigação de $100u.m.$, sujeitos a uma taxa de juro constante r, com uma taxa de recuperação R e um *spread* s. No caso de um empréstimo ou obrigação *default-free*, as $100u.m.$, na maturidade, converter-se-ão em $100(1+r)u.m.$.

Pode dizer-se que é possível obter uma relação semelhante à do Teorema 2.1, mas com uma medida martingala (pode ver-se um exemplo em [McNeil et al., 2005]). É suficiente definir a medida de probabilidade $\mathbb{Q} = (q, 1-q)$, de tal forma que:

$$\begin{aligned} B(0,1) &= \frac{100}{1+r} \\ \tilde{B}(0,1) &= \frac{1}{1+r}\Big((1-q)100 + R100q\Big) \end{aligned} \tag{2.8}$$

De acordo com as expressões (2.3) e (2.8) verifica-se que, o *spread* é dado por:

$$\begin{aligned} s(0,1) &= -\Big(ln\tilde{B}(0,1) - lnB(0,1)\Big) \\ &= -ln\left(\frac{100}{1+r}\Big((1-q)+Rq\Big)\right) + ln\left(\frac{100}{1+r}\right) \\ &= -ln\left(\frac{\frac{100}{1+r}\Big((1-q)+Rq\Big)}{\frac{100}{1+r}}\right) \\ &= -ln\left(1-(1-R)q\right) \end{aligned} \tag{2.9}$$

Proposição 2.2.
Para $(1-R)q << 1$*, tem-se que:*

$$s(0,1) \approx (1-R)q.$$

Esta relação entre o spread, a taxa de recuperação e a probabilidade de default é uma relação semelhante à obtida na Proposição 2.1 mas agora considerando a probabilidade de default dada na medida de probabilidade neutra face ao risco.

Demonstração. Considerando as séries geométricas,

$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$ e $\sum_{n=1}^{\infty} \frac{x^n}{n} = \int \frac{1}{1-x}dx = -ln(1-x)$, verifica-se que o erro cometido ao desprezar as potências de x de grau superior a 2 é pequeno se $x << 1$.

Portanto, pode escrever-se $-ln(1-x) \approx x$, e analogamente, $-ln\left(1-(1-R)q\right) \approx (1-R)q$. $\square$

2.2.2 Modelo em Tempo Contínuo

A Metodologia Actuarial

Considere-se novamente um espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$ e ainda um processo Browniano $(B_t)_{t\geq 0}$ sobre este espaço. Suponha-se também que existe um horizonte de tempo T e que o preço da obrigação com risco é dado por um processo estocástico $X_t^{\boldsymbol{\theta}}$, em função de um parâmetro $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ tal que, no caso da não ocorrência de *default*, se tem:

$$\begin{cases} dX_t^{\boldsymbol{\theta}} = \mu_t^{\boldsymbol{\theta}} dt + \sigma_t^{\boldsymbol{\theta}} dB_t, & t \in]0, T] \\ X_0^{\boldsymbol{\theta}} = X(0) \end{cases} \tag{2.10}$$

Pelo que, $(X_t^{\boldsymbol{\theta}})_{t\geq 0}$, o preço da obrigação sem incumprimento (*default*), é dado pela solução da Equação Diferencial Estocástica[1] (EDE) (2.10), em que $\mu_t^{\boldsymbol{\theta}}$ é o coeficiente de tendência (*drift*) e $\sigma_t^{\boldsymbol{\theta}}$ é o coeficiente de difusão (*volatility*). Os coeficientes de tendência e difusão podem depender de um parâmetro vectorial $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$. Por exemplo, no caso do modelo de Vasicek para a descrição da taxa de juro r (veja-se [Rutkowski e Musiela, 1998], pag.288), tem-se que:

$$dr_t = a(b - r_t)dt + \sigma dB_t,$$

logo $\boldsymbol{\theta} = (a, b, \sigma) \in (\mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \subseteq \mathbb{R}^3)$. Note-se que, neste modelo, o preço da obrigação de cupão zero é função de $(r_t),\ t > 0$.

Suponha-se agora que ocorre *default* no instante $\tau \in [0, T]$. Nesse caso, teremos a dinâmica do preço da obrigação dada por $(Y_t^{\boldsymbol{\theta}})_{t\in[0,T]}$, onde

$$Y_t^{\boldsymbol{\theta}} = X_t^{\boldsymbol{\theta}} \mathbb{1}_{[0,\tau[}(t) + X_t^{\boldsymbol{\theta}'} \mathbb{1}_{[\tau,T]}(t). \tag{2.11}$$

o que significa que antes de ocorrer *default* o processo evolui de acordo com a EDE (2.10) para o valor do parâmetro $\boldsymbol{\theta}$ e, após ocorrer *default*, respeita a mesma EDE, mas com outro parâmetro de valor $\boldsymbol{\theta}'$. No que se segue, assume-se o seguinte conjunto de hipóteses:

- **H1**: Para todo o $t \in [0, T]$, tem-se $X_t^{\boldsymbol{\theta}'} = R X_t^{\boldsymbol{\theta}}$, com R a taxa de recuperação.
- **H2**: Os parâmetros r e R são constantes no tempo; o *spread* instantâneo, $s(t)$ para $t \geq 0$, é determinístico.
- **H3**: O momento de *default* τ, é um tempo de paragem em relação à filtração Browniana, mas τ e $(B_t)_{t\geq 0}$ são independentes.
- **H4**: $\forall t \geq 0,\ \mathbb{E}[X_t^{\boldsymbol{\theta}}] \neq 0$.

Recorrendo ao princípio de que o preço de um título com risco é menor que o preço do mesmo título sem risco, observamos que, com $R < 1$ e $X_T^{\boldsymbol{\theta}} \leq Y_T^{\boldsymbol{\theta}}$, o *spread*, através

[1] Para os resultados fundamentais sobre este assunto pode consultar-se [Oksendal, 2007].

do factor acumulado $exp \int_0^t s(u)du$, corrige a discrepância entre os títulos com risco e sem risco e, assim, em termos de expectativas de mercado, com os preços descontados à data $t = 0$, deve ter-se:

$$\mathbb{E}^{\mathbb{P}}\left[\frac{X_t^{\boldsymbol{\theta}}}{exp(rt)}\right] = \mathbb{E}^{\mathbb{P}}\left[\frac{Y_t^{\boldsymbol{\theta}}\ exp \int_0^t s(u)du}{exp(rt)}\right]. \tag{2.12}$$

Note-se que a expressão (2.11) implica que:

$$\mathbb{1}_{[0,\tau[}(t) + c\mathbb{1}_{[\tau,T]}(t) = \begin{cases} 1 & \text{se } t < \tau \\ c & \text{se } t \geq \tau \end{cases}, \tag{2.13}$$

De acordo com as equações (2.11), (2.12) e (2.13), tem-se que:

$$\begin{aligned} exp\left(-\int_0^t s(u)du\right) \cdot \mathbb{E}^{\mathbb{P}}\left[\frac{X_t^{\boldsymbol{\theta}}}{exp(rt)}\right] &\overset{indep.}{=} \mathbb{E}^{\mathbb{P}}\left[\frac{X_t^{\boldsymbol{\theta}}}{exp(rt)}\right] \cdot \mathbb{E}^{\mathbb{P}}\left[\mathbb{1}_{[0,\tau[}(t) + R\mathbb{1}_{[\tau,T]}(t)\right] \Leftrightarrow \\ \Leftrightarrow exp\left(-\int_0^t s(u)du\right) &= \mathbb{E}^{\mathbb{P}}\left[\mathbb{1}_{[0,\tau[}(t) + R\mathbb{1}_{[\tau,T]}(t)\right] \\ &= \mathbb{P}\left[\tau > t\right] + R\mathbb{P}\left[\tau \leq t\right] \\ &= 1 - \mathbb{P}\left[\tau \leq t\right] + R\mathbb{P}\left[\tau \leq t\right] \\ &= 1 + (R-1)\mathbb{P}\left[\tau \leq t\right]. \end{aligned} \tag{2.14}$$

A expressão que resulta do Teorema 2.2, seguinte, representa o *spread* em função da taxa de recuperação e de $p := \mathbb{P}\left[\tau \leq t\right]$, a probabilidade de *default* na medida natural.

Teorema 2.2. *No modelo a tempo contínuo para o risco de crédito, o spread relaciona-se com a taxa de recuperação e a probabilidade de default através de:*

$$\int_0^t s(u)du = -\frac{1}{t}ln\Big(1 + (R-1)\mathbb{P}\left[\tau \leq t\right]\Big). \tag{2.15}$$

Demonstração. Dedução imediata a partir de (2.14). □

Proposição 2.3. *Escolhendo as unidades de forma a que na maturidade se tenha $t = 1$ e para $(1-R)p << 1$ na medida natural, tem-se que o valor médio do spread, $\overline{s}$, verifica:*

$$\overline{s} := \int_0^1 s(u)du \approx (1-R)p.$$

Demonstração. Análoga à demonstração da Proposição 2.2.

□

O resultado da Proposição 2.3 é semelhante ao resultado das Proposições 2.1 e 2.2.

A escolha das probabilidades de *default* (na medida natural ou na medida martingala) a usar, deve ser coerente, em particular, se uma necessita da *term structure* associada ao *spread*. Parece claro utilizar as probabilidades naturais (metodologia actuarial), se a intenção for estimar os parâmetros do modelo. As conexões com o modelo de Merton (ver [Rutkowski e Musiela, 1998])para risco de crédito devem ser analisadas posteriormente.

Para reforçar os resultados anteriores, apresentam-se seguidamente duas abordagens de diferentes fontes mas que convergem para um resultado idêntico ao obtido nesta secção.

2.3 Cálculo Aproximado do *Spread*

Nesta secção faz-se uma pequena ilustração do método aproximado para o cálculo do *spread* baseado numa matriz de transição qualquer, obtida a partir de dados históricos da carteira. Veja-se, por exemplo, a matriz de transição da Moody's na Tabela A.1, em anexo.

[Arvantis e Gregory, 2001], consideram que o preço, hoje, de uma *zero coupon bond* livre de risco e com maturidade T é $B(0,T)$. Inerente a uma taxa de juro composta continuamente, r, e para que não haja arbitragem, pode-se representar o preço desta *zero coupon bond* como em (2.16).

$$\begin{aligned} & B(0,T) \times exp(rT) = 1 \\ \Leftrightarrow \quad & B(0,T) = exp(-rT). \end{aligned} \tag{2.16}$$

Por outro lado, assume-se que o retorno de uma *zero coupon bond* com risco e com a mesma maturidade será $r+s$, onde s representa um *spread* de crédito, que é uma compensação para o risco de *default*. Neste caso, o preço desta obrigação, à data zero, deve ser menor que o de uma *zero coupon bond* com as mesmas características mas sem risco, e pode ser escrito como:

$$\tilde{B}(0,T) = k\ exp(-rT), \tag{2.17}$$

com $0 < k < 1$ e, portanto, pode dizer-se que $\exists s > 0$ tal que $k = exp(-sT)$. Logo a partir de (2.17), pode obter-se:

$$\tilde{B}(0,T) = exp\left[-(r+s)T\right]. \tag{2.18}$$

Considerando r fixo em $\mathbb{R}^+$ temos que, o preço de uma *zero coupon bond* com risco, $\tilde{B}(0,T)$, é igual ao valor esperado do preço P, onde este preço se pode definir através da função de probabilidade:

$$P = \begin{cases} R(T) & 1u.m. \\ p(0,T) & (1 - p(0,T)) \end{cases} \tag{2.19}$$

no caso de ocorrência ou não de *default*, respectivamente.

Portanto, é possível escrever o preço justo de uma *zero coupon bond* com risco, em termos de uma obrigação livre de risco, $\tilde{B}(0,T)$, da taxa de recuperação, $R(T)$, e da probabilidade de *default*, $p(0,T)$, no intervalo $[0,T]$ da seguinte forma:

$$\begin{aligned} \tilde{B}(0,T) &= \mathbb{E}[P] \Leftrightarrow \\ \Leftrightarrow \tilde{B}(0,T) &= B(0,T)\left[1 - p(0,T)\right] + R(T)B(0,T)p(0,T) \\ &= B(0,T)\left[1 - (1 - R(T))p(0,T)\right]. \end{aligned} \tag{2.20}$$

Então, a partir das equações anteriores, podemos obter o *spread* de crédito de uma obrigação de risco, em termos de probabilidade de *default* histórica, $p(0,T)$.

$$\begin{aligned} \frac{\tilde{B}(0,T)}{B(0,T)} &= \frac{exp\left[-(r+s)T\right]}{exp\left[-rT\right]} \\ &= exp\left[-sT\right] \\ &= \left[1 - (1 - R(T))p(0,T)\right]. \end{aligned} \tag{2.21}$$

O valor $p(0,T)$ pode ser calculado elevando a matriz de transição a um ano à potência apropriada. Note-se que é uma forma simplificada de extrair o *spread* baseado em dados históricos, mas não é, no entanto, a forma apropriada para apreçar uma obrigação de risco, uma vez que a probabilidade não é a probabilidade neutra face ao risco.

Proposição 2.4. *A equação* (2.21) *pode ser simplificada resultando:*

$$s \approx (1 - R)p. \tag{2.22}$$

Demonstração. Considerando o desenvolvimento de Taylor da função $f(x) = e^x$ com resto de grau 2 pode escrever-se:

$$exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n} \approx 1 + x; \quad x << 1$$

Então fazendo:

$$
\begin{aligned}
exp\,[-sT] &= [1-(1-R(T))p(0,T)] \\
1-sT &\approx [1-(1-R(T))p(0,T)] \\
s &\approx \frac{(1-R(T))p(0,T)}{T},
\end{aligned}
$$

para $T = 1$, resulta $s \approx (1-R)p$. □

O resultado da Proposição 2.4 coincide, mais uma vez, com os resultados anteriormente obtidos.

2.4 *Probabilidades de Default Implícitas de Mercado*

Uma outra abordagem possível para determinar o *spread* pode ser feita a partir de *corporate bonds.*

Estimação a partir de *Corporate Bonds*

Segundo [Pereira, 2009], tendo informação sobre os preços de mercado de *corporate bonds* é possível fazer *"reverse engineering"* para calcular as probabilidades que devem ser utilizadas para determinar os preços das mesmas *corporate bonds.* A estas probabilidades dá-se o nome de *probabilidades de default implícitas de mercado.*

O preço a 1-período de uma *zero coupon bond* com risco de crédito, pode ser determinado através de dois métodos equivalentes. Considerem-se as notações definidas anteriormente:

1. Sob o risco neutro (*risk-neutral*) de apreçamento, o valor actual $\tilde{B}(0,T)$ deve ser o *payoff* esperado (com nominal, $B(0,T)$) descontado à taxa sem risco (*risk-free rate*), r:

   $$\tilde{B}(0,T) = \frac{\mathbb{E}_0^Q[P]}{1+r} = \frac{(1-q(0,T))B(0,T) + q(0,T)R(T)B(0,T)}{1+r}, \qquad (2.23)$$

   onde $q(0,T)$ é a probabilidade de *default* .

   Pode-se designar por apreçamento com risco neutro pois o factor usual de correcção de risco, o *spread*, não consta no denominador. Em vez disso a aversão ao risco é descrita pelas probabilidades.

   Deste modo, q é a probabilidade de apreçamento ou de medida martingala. Esta probabilidade é superior à probabilidade de *default* histórica, p, pois inclui o prémio do risco de incumprimento ($q > p$).

2. Alternativamente, pode-se acrescentar o *spread*, s, no denominador:

$$\tilde{B}(0,T) = \frac{B(0,T))}{1+r+s}. \tag{2.24}$$

Utilizando as expressões (2.23) e (2.24) vem que:

$$\frac{(1-q)+qR}{1+r} = \frac{1}{1+r+s}$$

que pode ser resolvida de forma a obtermos q,

$$\begin{aligned} q &= \frac{s}{(1+r+s)(1-R)} \\ &\approx \frac{s}{1-R}. \end{aligned} \tag{2.25}$$

Em particular, a última aproximação diz que o *spread* é igual ao produto da probabilidade de *default* pela *loss given default*, ou seja,

$$s \approx q(1-R) = qL. \tag{2.26}$$

Isto significa que o *spread* é aproximadamente igual à perda esperada (com risco neutro).

As probabilidades de default implícitas de mercado são utilizadas para apreçamento de derivados de crédito. No entanto, não são utilizadas na gestão de risco de crédito (por exemplo, na determinação das reservas de capital), pois estão distorcidas pelo prémio de risco de crédito, ou seja, não são iguais às frequências esperadas de perda.

Foram apresentadas nesta secção diversas metodologias para obter uma expressão que definisse o *spread* em função da probabilidade de *default* e da taxa de recuperação. Embora com metodologias distintas, o resultado a que se chegou em cada uma delas foi idêntico, como tal, considera-se, neste trabalho, que o *spread* que se deve aplicar a um dado cliente solicitante de crédito pode ser dado pela expressão:

$$s \approx p(1-R), \tag{2.27}$$

em que p é a probabilidade de *default* na sua medida natural e R a taxa de recuperação.

No final do próximo capitulo, dada uma taxa de recuperação R, poder-se-á utilizar a equação (2.27) para determinar o *spread* que se deverá aplicar a alguns clientes solicitantes de crédito à Habitação, com base nas probabilidades de *default* que serão estimadas e numa taxa de recuperação constante.

Capítulo 3

Modelação de Risco de Crédito - Aplicação

3.1 A Carteira de Crédito

3.1.1 Definição de Cliente incumpridor

No desenvolvimento de um trabalho sobre risco de crédito, no qual se estudam modelos de *Credit Scoring*, é muito importante a definição de *Qualidade de Crédito*, isto é, o conceito de cliente cumpridor e cliente incumpridor utilizado neste trabalho, visto que a variável dependente no nosso modelo de avaliação de risco é a *Qualidade de Crédito*.

De uma forma geral, consideram-se clientes incumpridores aqueles clientes com um atraso no pagamento superior a 30 dias em pelo menos uma prestação do empréstimo. Os clientes cumpridores são aqueles que não possuem atrasos ou que possuem atrasos no máximo de 30 dias em alguma parcela do empréstimo. Por outras palavras, se o número de prestações já pagas for inferior ao número de prestações decorridas então o cliente está em incumprimento.

Podemos resumir as considerações da variável *Qualidade de Crédito*, no presente estudo, com a tabela 3.1.

Atraso no Pagamento (Dias)	Situação
De 0 a 30 dias	Cumpridor
Mais de 30 dias	Incumpridor
Pagas $<$ Decorridas	Incumpridor

Tabela 3.1: Definição de Qualidade de Crédito

3.1.2 Composição da Carteira de Crédito

De forma a modelar o risco de crédito de uma carteira de clientes, recorreu-se a dados históricos de uma carteira de empréstimos bancários para créditos à Habitação.

A carteira de crédito utilizada neste trabalho consiste numa carteira de crédito à Habitação e é composta por 12.000 clientes dos quais cerca de 3% entraram em incumprimento.

Devido à falta de dados reais tornou-se necessário simular os valores de todas as variáveis que considerámos de interesse para o estudo. A simulação da carteira de crédito fez uso de métodos de aleatoriedade, e foi obtida recorrendo à aplicação *Microsoft Excel*. No presente estudo, a variável de interesse é a *Qualidade de Crédito* dos clientes, que consiste numa variável nominal, em que são atribuídos os valores 0 para clientes cumpridores e 1 para clientes incumpridores. Após a simulação de todas as variáveis para os 12000 clientes, agruparam-se os dados de forma a termos uma base de dados pronta para utilizar no desenvolvimento do modelo de *Credit Scoring*. Pode-se visualizar na Tabela A.3 em anexo, uma parte da carteira de crédito simulada.

Para o tratamento dos dados da carteira de crédito dividiu-se a mesma em duas amostras, a amostra de desenvolvimento e a amostra de validação. A amostra de desenvolvimento é assim constituída por 10000 dos clientes do total da carteira de crédito e a amostra de teste é constituída pelos restantes 2000 clientes. A amostra de desenvolvimento é a amostra que será utilizada de agora em diante para a estimação do modelo que melhor explica a variável dependente em função das variáveis explicativas. A amostra de teste apenas será utilizada para avaliar a capacidade preditiva do modelo ajustado com a amostra de desenvolvimento.

Tratamento de dados

O tratamento de dados numa amostra de uma carteira de crédito real consistiria, com base nos dados recolhidos, na identificação das variáveis estatisticamente significativas e no desenvolvimento dos modelos de *Credit Scoring*.

No entanto, neste trabalho, foram escolhidas, com base em estudos anteriores, as variáveis que se consideraram importantes recolher aquando um novo pedido de crédito. Posteriormente, foram simulados os dados correspondentes a todas as variáveis e obtendo-se, desta forma, a base de dados que servirá de base a este trabalho. Finalmente, com a base de dados construída, aplicou-se um dos métodos estatísticos utilizados em modelos de Credit Scoring, a Regressão Logística, e foram efectuados estudos estatísticos para identificar as variáveis mais significativas que permitem "explicar" a Qualidade de Crédito de um cliente.

3.1.3 As Variáveis Utilizadas

Nos modelos de *Credit Scoring*, a variável resposta (dependente) é a *Qualidade de Crédito* (*default* ou *non-default*). Para classificar as observações de acordo com a Qualidade de Crédito, foram seleccionadas variáveis explicativas, ou independentes, que pudessem influenciar a situação dos clientes nas suas obrigações enquanto credores.

Para melhor orientação na definição e escolha das variáveis explicativas foram também consultados outros trabalhos de *Credit Scoring*, entre eles, um trabalho elaborado pela CrediRisk para um banco português, o trabalho de dissertação de [Araújo, 2006] e a página Web [Wikipédia, 2009b].

Inicialmente, foi construída uma base de dados que agrega um conjunto de possíveis variáveis explicativas pré-seleccionadas para utilização na construção dos modelos. A partir desta base inicial, foram seleccionadas, através da aplicação de métodos estatísticos, as variáveis explicativas mais relevantes, as quais efectivamente compuseram os modelos.

A Tabela 3.2 ilustra as variáveis explicativas adoptadas bem como algumas características das mesmas.

	Variáveis	Código da Variável	Natureza da Variável	Tipo	Simulação
1	Idade	Idade	Qualitativa	Categórica	Distribuição Uniforme(a,b)
2	Nº de Dependentes	Dep	Qualitativa	Categórica	Distribuição Uniforme(a,b)
3	Tipo Habitação	Hab	Qualitativa	Categórica	Distribuição Uniforme(a,b)
4	Idade casa	Idcasa	Qualitativa	Categórica	Distribuição Uniforme(a,b)
5	Estado Civil	Civil	Qualitativa	Categórica	Distribuição Uniforme(a,b)
6	Código Postal	Postal	Qualitativa	Categórica	Distribuição Uniforme(a,b)
7	Tipo Profissão	Prof	Qualitativa	Categórica	Distribuição Uniforme(a,b)
8	Antiguidade Profissão	Antg	Qualitativa	Categórica	Distribuição Uniforme(a,b)
9	Rendimento Líquido	Rend	Qualitativa	Categórica	Distribuição Uniforme(a,b)
10	Crédito	Cred	Quantitativa	Numérica	Distribuição Gamma(α;β)
11	% de Entrada	Entr	Qualitativa	Categórica	Distribuição Beta(α;β)
12	NºTotal Prestações	Ptotal	Quantitativa	Numérica	Distribuição Uniforme(a,b)
13	NºPrestações Pagas	Pagas	Quantitativa	Numérica	Aleatório $[Min, Ptotal]$
14	Pagas/Total (%)	Pagpercent	Qualitativa	Categórica	-

Tabela 3.2: Lista Inicial de Variáveis Explicativas

A Tabela A.2 em anexo, ilustra as variáveis explicativas e as respectivas categorias utilizadas.

As variáveis categóricas também são designadas por variáveis *dummy*, dicotómicas ou binárias e são variáveis construídas artificialmente para medir a presença ou ausência de um determinado atributo.

A variável dependente, *Qualidade de Crédito*, também foi inserida na base de dados através do uso de variáveis *dummy*. Assim foi criada uma *dummy*, denominada "Default", que assume o valor 0 para clientes cumpridores e 1 para clientes incumpridores, visando diferenciar os clientes em relação à presença ou ausência da condição de *default*.

3.1.4 Simulação da Carteira de Crédito

Nesta secção será descrita a metodologia utilizada para gerar cada uma das variáveis constituintes da carteira de crédito. Para as variáveis categóricas, foram efectuados alguns pressupostos sobre o peso relativo de cada uma das categorias da variável, ou seja, a percentagem de clientes com essa característica, de forma a que a simulação resultasse numa carteira de clientes minimamente realista.

Variáveis Independentes

As covariáveis simuladas são:

Idade (Idade):
Para a variável *Idade* foram consideradas três categorias, conforme a Tabela A.2, em anexo.

Para cada cliente foi gerado um número aleatório entre 0 e 1 de acordo com a distribuição Uniforme. O cliente é afecto à categoria correspondente ao número aleatório gerado. Por exemplo se o número aleatório obtido se encontra entre 0 e 0.13 o cliente é inserido na categoria "Jovens".

A grande maioria das variáveis da Tabela A.2 foram construídas de forma análoga à variável *Idade*.

Nº de Dependentes (Dep):
A variável *Nºde Dependentes* foi simulada da mesma forma que a variável *Idade* tendo em conta as percentagens das respectivas categorias. No entanto, nesta variável, considera-se que os clientes com menos de 25 anos não têm quatro ou mais filhos, ou seja, fez-se depender esta variável da variável *Idade*, mas apenas para a categoria "Muitos".

Tipo Habitação (Hab):
A variável *Tipo Habitação* foi também simulada da mesma forma que a variável *Idade*, tendo em consideração as percentagens das respectivas categorias.

Idade Casa (Idcasa):
A variável *Idade Casa* foi simulada de forma análoga, tendo em conta as percentagens das respectivas categorias.

Estado Civil (Civil)
A variável *Estado Civil* foi simulada da mesma forma que a variável *Idade*, mas com as percentagens respectivas. Neste caso, em particular, as percentagens foram retiradas dos censos feitos em Portugal no ano de 2001 (veja-se [INE, 2002]).

Código Postal (Postal):
A variável *Código Postal* foi simulada da mesma forma que a variável *Idade*, tendo em conta as respectivas percentagens. No entanto, os grupos formados tiveram como base a informação do estudo da *CrediRisk* para o Crédito Automóvel. Por defeito,

consideramos as mesmas categorias, visto não termos dados reais para fazer um estudo estatístico sobre as zonas que se deveriam agrupar.

Os valores que constam na descrição da covariável *Código Postal* na Tabela A.2, correspondem as dois primeiros números dos códigos postais nacionais.

Tipo Profissão (Prof):

Esta variável foi simulada exactamente nas mesmas condições da variável *Código Postal*, utilizando também a informação do estudo da *CrediRisk*.

Antiguidade Profissão (Antg):

Para a variável *Antiguidade Profissão* consideraram-se apenas dois grupos e a simulação foi feita como anteriormente para a variável *Idade*.

Rendimento Líquido (Rend):

Mais uma vez, esta variável foi simulada tendo em conta as percentagens que foram atribuídas a cada categoria. Não se consideram os clientes com ordenados inferiores a €450 e superiores a €20000, ou seja, na simulação dos valores dos ordenados propriamente ditos, apenas se gerou um número aleatório entre 450 e 20000.

Crédito (Cred):

O procedimento adoptado na simulação da variável *Crédito* é bastante distinto das restantes variáveis. Por um lado, não é uma variável categórica e por outro, tem características muito próprias.

Considerou-se que a variável *Crédito* poderia ser modelada por uma distribuição $Y = aX + b$, onde $X \sim Gama(\alpha = 1.5; \beta = 3)$, $a = 20000$ e $b = 75000$. Os parâmetros foram escolhidos de forma a obter montantes de crédito compatíveis com uma carteira de crédito à habitação realista. Note-se que o montante médio de crédito concedido é aproximadamente €165.000 .

Percentagem de Entrada (Entr):

Para simular esta variável considerou-se que a *Percentagem de entrada* poderia ser modelada através de uma distribuição $Beta(\alpha; \beta)$. Os parâmetros foram escolhidos assumindo um valor mínimo de percentagem de entrada igual a 0, um valor máximo igual a 40% e consideranto o valor 5% como o valor mais provável. Assim considerou-se que Percentagem de Entrada $\sim Beta(\alpha = 1.4375; \beta = 4.3125)$

Nº Total Prestações (Ptotal):

Na simulação desta variável, não tendo disponíveis quaisquer estudos semelhantes para estimação da percentagem de clientes em cada categoria, começou-se por simular a percentagem de clientes, em cada uma das dez categorias definidas, utilizando para tal uma distribuição $Binomial(n = 10, p = 0.6)$. A escolha do parâmetro p reflecte, novamente, a expectativa de simular uma carteira realista. Uma vez estimadas as percentagens de clientes em cada classe, procedeu-se à simulação tal como, por exemplo, a variável *Idade*.

Nº Prestações Pagas (Pagas):

A simulação desta variável foi realizada gerando um número aleatório, uniformemente distribuído, entre 60 prestações (assumiu-se que os incumprimentos ocorrem apenas após 5 anos de crédito) e o *Nº Total de Prestações* do respectivo cliente.

Nº Prestações Pagas/Total (%)(Pagpercent):

Esta variável não é mais que o quociente entre o número de prestações pagas e o total de prestações para o mesmo cliente.

Variável Dependente

Default:

A variável dependente, neste modelo, será, como referido anteriormente, a Qualidade de Crédito de cada cliente, medida como a ocorrência ou não de *Default* ou incumprimento.

Como referido anteriormente, a metodologia utilizada neste trabalho pressupõe a análise de dados históricos da carteira, de forma a poder inferir sobre as variáveis explicativas de interesse no fenómeno de ocorrência de incumprimento.

Desta forma, tornou-se também necessário simular a variável dependente, para todos os clientes que compõem a carteira de crédito considerada.

Na simulação desta variável optou-se por fazer uma *"reverse engineering"*, ou seja, fazer alguns pressupostos sobre quais as covariáveis que, regra geral, terão mais influência na ocorrência de *default*. Esta opção prendeu-se, essencialmente, com o objectivo de introduzir alguma ordem na aleatoriedade das covariáveis, uma vez que foram obtidas por simulação e não reflectem as interacções que, em situações normais, quase certamente se verificarão.

A escolha das covariáveis consideradas com mais relevância na explicação da ocorrência de *default* baseou-se nos estudos anteriores já referidos, e na sensibilidade a este assunto reportada por especialistas da área de crédito.

A cada covariável foi atribuído um peso, conforme a sua importância na ocorrência de *default*. Nas variáveis categóricas, atribuiu-se também um peso a cada categoria, definindo assim, quais as categorias mais propensas à ocorrência de *default*.

Os pressupostos estabelecidos no peso de cada covariável e em cada uma das categorias encontram-se ilustrados na Tabela 3.3.

A simulação da variável *Default* depende da combinação das ponderações das covariáveis de cada cliente. A ponderação, para cada cliente, obtém-se da seguinte forma:

$$\sum_{i=1}^{n} pv_i \times pc_i, \tag{3.1}$$

Nome Variável	Peso Variável	Peso Categorias 1	2	3	4	5	6	7	8	9	10
Idade	0,1000	1	0,66	0,33							
Nº de Dependentes	0,0500	1	0,33	0,66							
Tipo Habitação	0,0050	1	0,33	0,66							
Idade casa	0,0050	1	0,66	0,33							
Estado Civil	0,0400	1	0,66	0,33							
Código Postal	0,1000	0,25	0,5	0,75	1						
Tipo Profissão	0,0095	1	0,75	0,5	0,25						
Antiguidade Profissão	0,0005	1	0,5								
Rendimento Líquido	0,2400	1	0,75	0,5	0,25						
Crédito	0,0500	1	0,75	0,5	0,25						
% de Entrada	0,1000	1	0,5								
NºTotal Prestações	0,1000	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
NºPrestações Pagas	0,2000	1	0,75	0,5	0,25						
	1,0000										

Tabela 3.3: Tabela de Pesos dos Factores Chave

onde $i = 1, ..., n$ são as covariáveis, pv_i é o peso da covariável i e pc_i é o peso da categoria da covariável i onde o cliente se insere.

A ponderação será um valor entre 0 e 1, uma vez que todos os pesos se encontram nesse intervalo. Após o cálculo da ponderação para cada cliente define-se se o cliente entrou ou não em incumprimento. Para tal, definiu-se um limite a partir do qual se consideraria que o cliente teria originado *default*. O valor mínimo de ponderação para a ocorrência de *default* foi escolhido de forma a garantir que a percentagem de clientes incumpridores fosse aproximadamente 3% dos clientes da carteira de crédito. Para a carteira de crédito simulada neste trabalho, o limite utilizado foi de 0.85.

3.2 Modelo e Técnicas Estatísticas Utilizadas

A construção de modelos quantitativos de *Credit Scoring*, a fim de auxiliar o processo de concessão de crédito à Habitação e avaliação de risco numa instituição bancária Portuguesa, é a ideia fulcral deste trabalho.

Nos modelos de *Credit Scoring* o objectivo principal é identificar os factores que determinam a probabilidade de incumprimento (*default*) dos clientes. Um dos pressupostos principais é admitir-se que as características dos clientes que entrarão em incumprimento no futuro são semelhantes às características daqueles que entraram em incumprimento no passado.

Nos modelos de aprovação de crédito utilizam-se informações existentes sobre o cliente. Estes modelos são essenciais como complemento na avaliação e na tomada de decisão de um analista aquando de um novo pedido de crédito.

3.2.1 Análise Estatística das Variáveis

Utilizando o *software* R[1] podemos tratar estatisticamente a base de dados da carteira de crédito. Começando por uma análise preliminar das estatísticas descritivas e depois por uma análise gráfica é possível ter uma visão mais geral das variáveis que constituem a carteira de crédito e das relações entre as mesmas.

Inicialmente carrega-se o conjunto de dados e realiza-se uma análise das estatísticas descritivas da qual resulta a Tabela 3.4.

Idade			Dep			Hab		
Categoria	**Freq.**	**%**	**Categoria**	**Freq.**	**%**	**Categoria**	**Freq.**	**%**
Adultos	8566	85.66%	Alguns	1540	15.4%	Casa	8822	88.22%
Idosos	207	2.07%	Muitos	460	4.6%	Comercio	830	8.3%
Jovens	1227	12.27%	Poucos	8000	80,00%	Empresa	348	3.48%
Idcasa			**Civil**			**Postal**		
Categoria	**Freq.**	**%**	**Categoria**	**Freq.**	**%**	**Categoria**	**Freq.**	**%**
Nova	7474	74.74%	Cas	4935	49.35%	Zona1	453	4.53%
Usada	2060	20.6%	Div	262	2.62%	Zona2	1014	10.14%
Velha	466	4.66%	Sol	4803	48.03%	Zona3	2064	20.64%
						Zona4	6469	64.69%
Prof			**Antg**			**Rend**		
Categoria	**Freq.**	**%**	**Categoria**	**Freq.**	**%**	**Categoria**	**Freq.**	**%**
Tipo1	1956	19.56%	Antigo	9295	92.95%	Alto	302	3.02%
Tipo2	2968	29.68%	Recente	705	7.05%	Baixo	3699	36.99%
Tipo3	3616	36.16%				Medio	1996	19.96%
Tipo4	1460	14.6%				Minimo	4003	40.03%
Cred			**Entr**			**Ptotal**		
Medida	**Localização**		**Categoria**	**Freq.**	**%**	**Medida**	**Localização**	
Min.	75163		Mais10	4393	43.93%	Min.	60.0	
1st Qu.	111440		Menos10	5607	56.07%	1st Qu.	300.0	
Median	146903					Median	360.0	
Mean	165521					Mean	361.9	
3rd Qu.	199062					3rd Qu.	420.0	
Max.	661874					Max.	600.0	
Pagas			**Pagpercnt**			**Default**		
Medida	**Localização**		**Categoria**	**Freq.**	**%**	**Categoria**	**Freq.**	**%**
Min.	60.0		Entre25e50	2887	28.87%	0	9673	96.73%
1st Qu.	124.0		Entre50e75	3092	30.92%	1	327	3.27%
Median	194.0		Mais75	3055	30.55%			
Mean	208.6		Menos25	966	9.66%			
3rd Qu.	279.0							
Max.	578.0							

Tabela 3.4: Análise Preliminar das Variáveis

Uma outra forma de analisar os dados da Tabela 3.4 é o recurso a histogramas e a diagramas "caixa-e-bigodes".

[1] The R Project for Statistical Computing - http://www.r-project.org/

Para ilustrar os dados das variáveis qualitativas da Tabela 3.4, realizaram-se os histogramas da Figura 3.1.

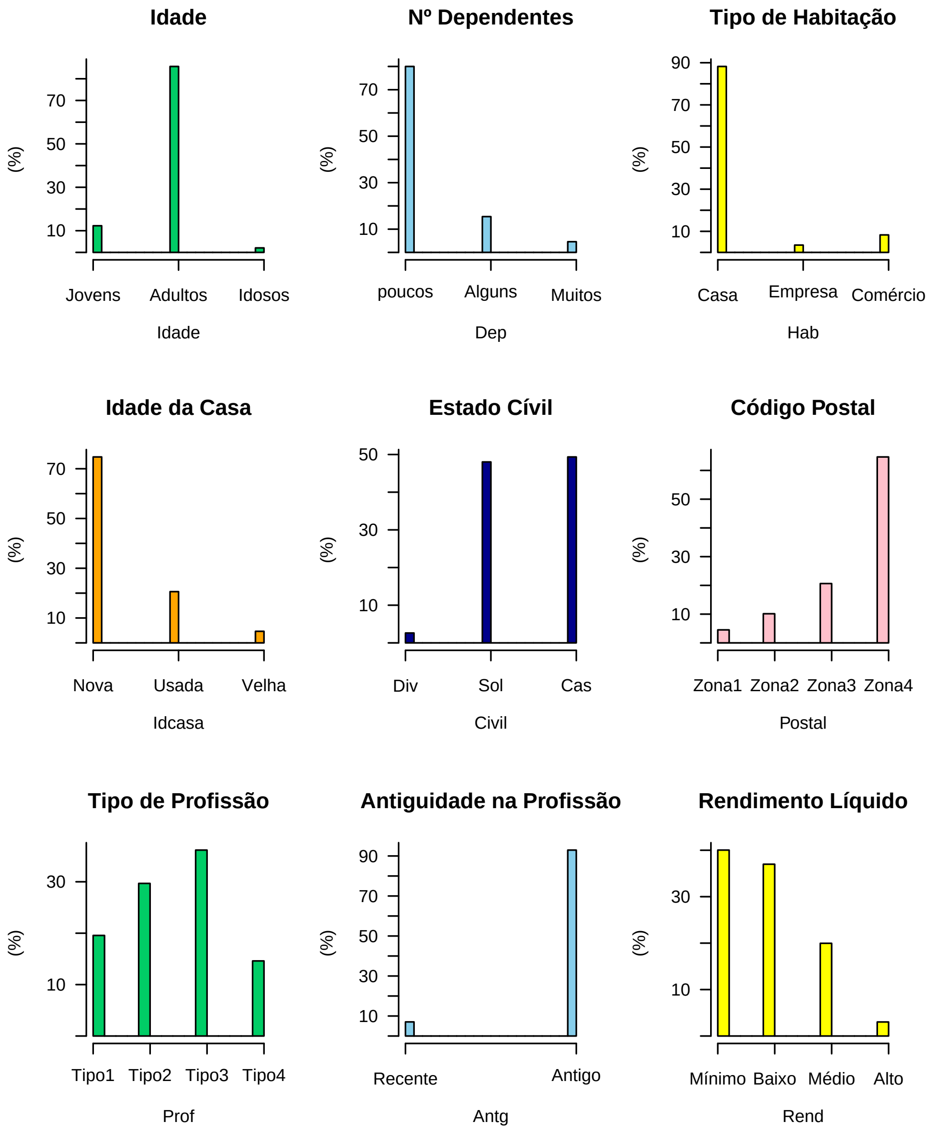

Figura 3.1: Histogramas Variáveis Qualitativas

Realizaram-se também, alguns gráficos com os dados da carteira de crédito, para ilustrar a relação entre as covariáveis e a variável resposta, *"Default"*. Como as variáveis explicativas são variáveis categóricas não é possível fazer uma análise através de gráficos de dispersão, então utilizam-se os gráficos "caixa-e-bigodes".

Segundo [Murteira et al., 2002], o rectângulo ("a caixa") é desenhado de tal modo, que os seus topos inferior e superior correspondem aos 1º e 3º quartis. O segmento que divide a caixa em duas partes corresponde à mediana e é possível identificar as assimetrias nos dados, caso as partes, em que a caixa está dividida, sejam muito diferentes. A caixa nos seus limites horizontais, isto é, entre o primeiro e o terceiro quartis, contém 50% das observações. O mínimo e o máximo são representados pelos segmentos inferior e superior ("os bigodes") desenhados no exterior do rectângulo e servem para identificar a existência de *outliers*, no caso de os bigodes serem, relativamente à caixa, muito grandes.

Os *outliers* são identificados da seguinte forma:

1. x_i é um *outlier* severo se:
$$x_i < Q_1 - 3(Q_3 - Q_1) \quad ou \quad x_i > Q_3 + 3(Q_3 - Q_1);$$

2. x_i é um *outlier* moderado se:
$$\begin{aligned} & Q_1 - 3(Q_3 - Q_1) < x_i < Q_1 - 1.5(Q_3 - Q_1) \\ ou \quad & Q_3 + 1.5(Q_3 - Q_1) < x_i < Q_3 + 3(Q_3 - Q_1). \end{aligned}$$

A Figura 3.2 é a representação gráfica da variável *Crédito*. Neste gráfico, encontram-se representados os valores que constam na Tabela 3.4 relativamente a esta variável.

Como se pode observar, existe uma notória assimetria entre os montantes de crédito e verifica-se, a existência de *outliers* moderados e severos, pois o tamanho do bigode superior é superior a uma vez e meia a distância entre quartis e superior a 3 vezes a mesma distância, respectivamente.

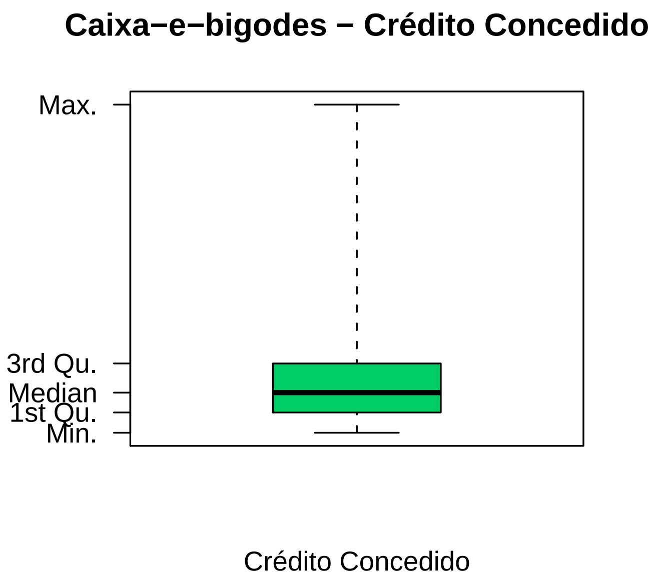

Figura 3.2: Caixa-e-Bigodes: Variável *Crédito*

A Figura 3.3 ilustra as caixas de bigodes para as *Prestações Totais* e *Prestações Pagas*. Para ambas as variáveis é possível observar-se a existência de *outliers* moderados mas apenas para a variável *Prestações Pagas* se verifica uma ligeira assimetria entre os dados.

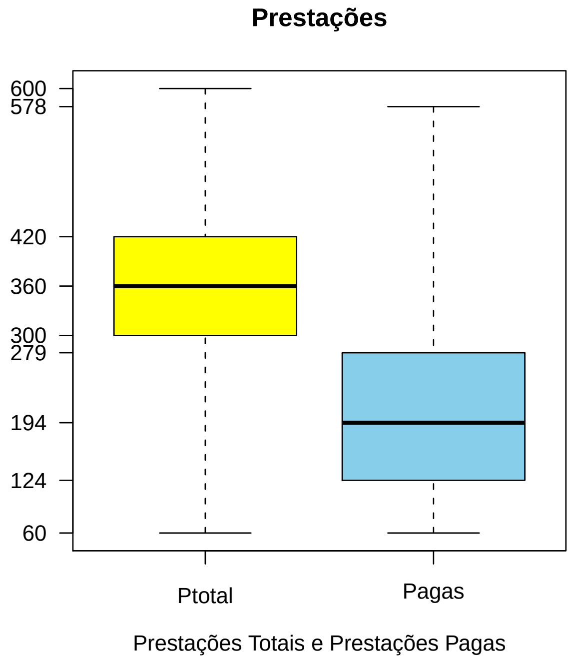

Figura 3.3: Caixa-e-Bigodes: *Prestações Totais* e *Prestações Pagas*

Os histogramas da Figura 3.4 representam a frequência relativa das variáveis quantitativas, *Crédito* e *Prestações Totais* e *Prestações Pagas*.

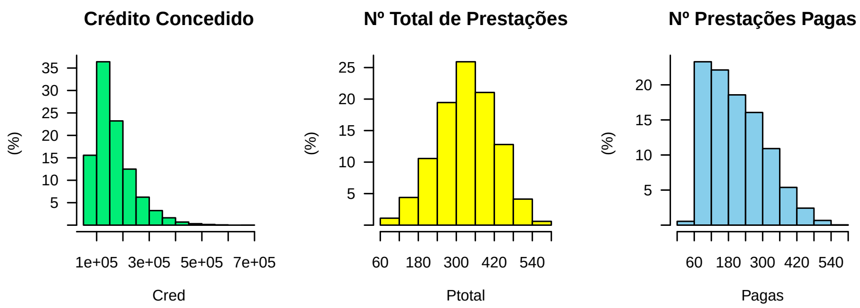

Figura 3.4: Histogramas Variáveis Quantitativas

Tal como nas Figuras 3.2 e 3.3, o histogramas da variável *Crédito* e da variável *Prestações Pagas* mostram a assimetria existente entre os os montantes de crédito e entre o número de prestações pagas, respectivamente. Observa-se também a ligeira simetria entre os valores das prestações totais, que se justifica pela forma como esta variável foi simulada.

Para analisar a relação entre a variável resposta *Default* e as variáveis quantitativas *Credito*, *Prestações Totais* e *Prestações Pagas* construíram-se os gráficos da Figura 3.5 e da Figura 3.6.

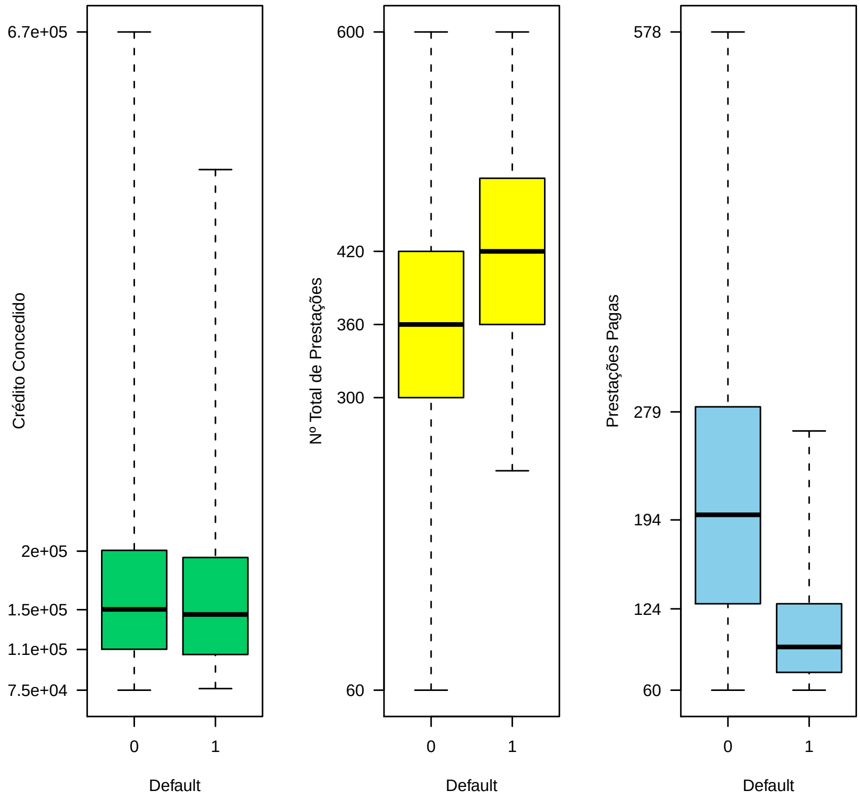

Figura 3.5: Caixa-e-Bigodes: Variável *Default* vs Prestações

Note-se que na Figura 3.5, no gráfico "Default/Crédito Concedido", os clientes que entram em *default*, regra geral, são clientes cujos créditos são montantes mais baixos, existem poucos clientes com crédito superior a €300.000 que entram em incumprimento.

Em analogia com o gráfico respectivo da Figura 3.6, as observações são semelhantes, existem alguns clientes com créditos elevados (*outliers*) que entram em incumprimento provocando os dois picos a verde escuro, mas os restantes incumprimentos situam-se numa zona onde o montante do crédito é inferior.

Observando agora os gráficos *Default/Pagas* das mesmas figuras, verifica-se que a ocorrência de incumprimento ocorre maioritariamente no início dos contratos. Compa-

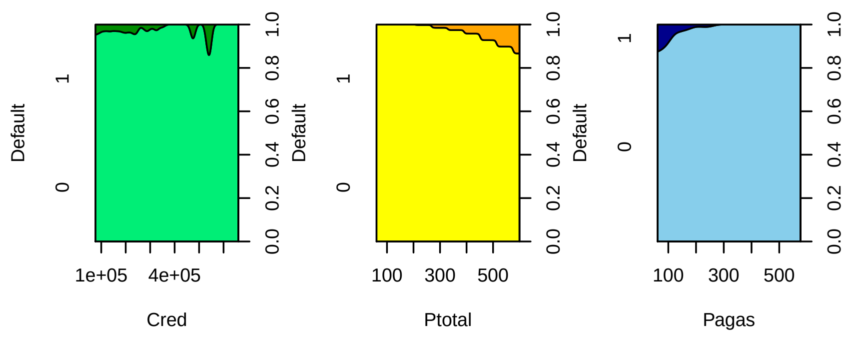

Figura 3.6: CdPlot - Relação entre Default e as Variáveis Quantitativas

rando os gráficos com os valores da Tabela 3.3, mais uma vez se verifica que os gráficos reflectem os pesos dados às variáveis. Neste caso em particular, foi dado peso 1 à menor percentagem de prestações pagas e 0.25 quando as prestações pagas ultrapassam 75% do número total de prestações. Estes pesos foram dados, com base na ideia de que, alguém que ainda só pagou algumas prestações mais facilmente abandona o compromisso de crédito, em caso de dificuldade, pois não perde tanto como se já estivesse no fim do empréstimo.

Para o *Nº Total de Prestações* verifica-se uma situação oposta, ou seja, quanto maior o *Nº Total de Prestações* em que o crédito pode ser pago, maior a probabilidade de ocorrer Default, e novamente estas observações são justificáveis pois quanto maior o *Nº Total de Prestações* maior é o tempo disponível para ocorrerem eventos que possam levar ao incumprimento, tais como desemprego ou doença, entre outros. Novamente, os gráficos são o reflexo dos pesos atribuídos às variáveis (rever Tabela 3.3).

Uma outra forma de analisar a variável *Default* em relação às *Prestações Pagas* é o recurso aos *scatterplots* dos dados, desenhando uma linha através dos mesmos. Esta linha é desenhada através de uma função que utiliza informação proveniente do modelo ajustado para produzir "funções suaves". Ajustaram-se os pares de variáveis *Default vs Crédito* e *Default vs Prestações Pagas* e obtiveram-se os gráficos da Figura 3.7.

Observa-se que a curva de ajustamento da variável *Crédito*, em comparação com a curva de ajustamento da variável *Prestações Pagas*, está muito mais próxima dos valores correspondentes ao *Default* igual a zero. Pode dizer-se que, embora a variável *Crédito* seja significativa para determinar o ocorrência ou não de incumprimento, a variável *Prestações Pagas* é mais significativa. O declive mais acentuado da curva de ajustamento da variável *Prestações Pagas* sugere que esta variável tenha mais peso na ocorrência de *default*. A variável *Crédito* parece explicar melhor a ocorrência de cumprimento do que a ocorrência de incumprimento.

A Tabela 3.5 faz um resumo da frequência absoluta e da frequência relativa com que

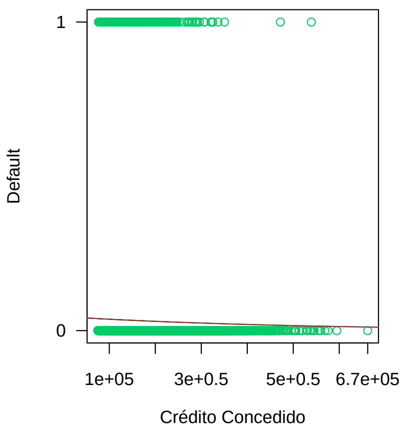

Figura 3.7: Ajustamento Usando a Função Predict

cada variável qualitativa e respectiva categoria ocorrem diferenciando ainda o número de clientes que entrou em incumprimento e os que cumpriram o contrato.

Concluída a análise preliminar dos dados da carteira de crédito é importante descrever os métodos e modelos estatísticos que serão utilizados para estudar de forma mais intensiva os dados e, desta forma, determinar quais as variáveis explicativas mais significativas para explicar a ocorrência de incumprimento para os clientes desta carteira.

Quando o objectivo é estudar a relação entre variáveis ou mais especificamente, analisar a influência que uma variável ou mais variáveis (*explicativas*), medidas em indivíduos ou objectos, têm sobre uma variável de interesse (*variável resposta*), geralmente o problema é abordado através do estudo de um modelo de regressão que relacione essa variável de interesse com as variáveis explicativas. Por esta razão, neste trabalho serão utilizados os Modelos Lineares Generalizados.

Variável	Categoria	0 Frq. Abs.	0 Frq. Rel.	1 Frq. Abs.	1 Frq. Rel.
Idade	Adultos	8330	83,30%	236	2,36%
	Idosos	204	2,04%	3	0,03%
	Jovens	1139	11,39%	88	0,88%
Dep	Alguns	1514	15,14%	26	0,26%
	Muitos	453	4,53%	7	0,07%
	Poucos	7706	77,06%	294	2,94%
Hab	Casa	8538	85,38%	284	2,84%
	Comercio	798	7,98%	32	0,32%
	Empresa	337	3,37%	11	0,11%
Idcasa	Nova	7240	72,40%	234	2,34%
	Usada	1983	19,83%	77	0,77%
	Velha	450	4,50%	16	0,16%
Civil	Cas	4822	48,22%	113	1,13%
	Div	247	2,47%	15	0,15%
	Sol	4604	46,04%	199	1,99%
Postal	Zona1	452	4,52%	1	0,01%
	Zona2	1005	10,05%	9	0,09%
	Zona3	2013	20,13%	51	0,51%
	Zona4	6203	62,03%	266	2,66%
Prof	Tipo1	1886	18,86%	70	0,70%
	Tipo2	2870	28,70%	98	0,98%
	Tipo3	3501	35,01%	115	1,15%
	Tipo4	1416	14,16%	44	0,44%
Antg	Antigo	8990	89,90%	305	3,05%
	Recente	683	6,83%	22	0,22%
Rend	Alto	302	3,02%	0	0,00%
	Baixo	3679	36,79%	20	0,20%
	Medio	1995	19,95%	1	0,01%
	Minimo	3697	36,97%	306	3,06%
Entr	Mais10	4359	43,59%	34	0,34%
	Menos10	5314	53,14%	293	2,93%
Ptotal	60	12	0,12%	0	0,00%
	120	98	0,98%	0	0,00%
	180	439	4,39%	0	0,00%
	240	1055	10,55%	2	0,02%
	300	1916	19,16%	29	0,29%
	360	2526	25,26%	66	0,66%
	420	2017	20,17%	88	0,88%
	480	1187	11,87%	92	0,92%
	540	371	3,71%	42	0,42%
	600	52	0,52%	8	0,08%

Tabela 3.5: Default vs Variáveis Qualitativas

3.2.2 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG) vêm unificar modelos anteriores desenvolvidos para a modelação estatística e foram introduzidos por [Nelder e Wedderburn, 1972]. Nos MLG's a variável resposta segue uma distribuição dentro de uma família de distribuições com propriedades muito específicas: a **Família Exponencial**.

Os modelos que se enumeram a seguir são caso particulares dos modelos lineares generalizados:

- modelo de regressão linear clássico,
- modelos de análise de variância e covariância,

- modelo de regressão logística,
- modelo de regressão de Poisson,
- modelos log-lineares para tabelas de contingência multidimensionais,
- modelo *probit* para estudos de proporções, etc

Notação e Terminologia

Considere-se, ver [Turkman e Silva, 2000], a variável aleatória Y, de interesse primário, e que designa *variável resposta* ou *variável dependente*, e um vector $\mathbf{x} = (x_1, \ldots, x_k)^T$ de k variáveis explicativas, também designadas por *covariáveis* ou *variáveis independentes*, que se crê explicarem parte da variabilidade inerente a Y.
A variável resposta Y pode ser contínua, discreta ou dicotómica. As covariáveis, determinísticas ou estocásticas, podem ser também de qualquer natureza: contínuas, discretas, qualitativas de natureza ordinal ou dicotómicas. Assume-se que os dados têm a forma

$$(y_i, x_i), \quad i = 1, \ldots, n, \tag{3.2}$$

resultantes da realização de $(Y, \mathbf{x})$ em n indivíduos ou unidades experimentais, sendo as componentes Y_i do vector aleatório $\mathbf{Y} = (Y_i, \ldots, Y_n)^T$ independentes. Pode-se representar (3.2) na forma matricial,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} x_{11} & x_{11} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}. \tag{3.3}$$

Os modelos lineares generalizados são uma extensão do modelo linear clássico,

$$\begin{aligned} \mathbf{Y} &= \mathbf{Z}\boldsymbol{\beta} + \varepsilon \\ &\mathit{ou} \\ \mathbf{Y} &= \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + \boldsymbol{\varepsilon}. \end{aligned} \tag{3.4}$$

onde $\mathbf{Z}$ é uma matriz de dimensão $n \times p$ de especificação do modelo, associada a um vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ de parâmetros e $\boldsymbol{\varepsilon}$ um vector de erros aleatórios.

A escolha da função de ligação depende do tipo de resposta e do estudo particular que se está a fazer. Por exemplo, para dados binários utiliza-se a função de ligação *Logit* que será vista mais adiante na explicação do modelo de regressão Logística.

Família Exponencial

Definição 3.1. ***Família Exponencial***
Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever na forma:

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta-b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

onde θ e ϕ são parâmetros escalares, $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas.

Neste trabalho considera-se $\phi = 1$.

Para aplicar a metodologia dos MLG's a um conjunto de dados há necessidade, após a formulação do modelo que se pensa adequado, de proceder à realização de inferências sobre esse modelo. A inferência em MLG's baseia-se essencialmente na verosimilhança.

Testes de Hipóteses

Quando estamos perante um problema de selecção de covariáveis e queremos testar se um submodelo é melhor que o modelo original é comum utilizar a *Estatística de Wald*, a *Estatística de Wilks* ou *Estatística de Razão de Verosimilhanças*, ver [Turkman e Silva, 2000]. Estas estatísticas são deduzidas a partir das distribuições assimptóticas dos estimadores de máxima verosimilhança e de funções adequadas desses estimadores.

Considera-se o teste de hipóteses da forma:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi} \quad versus \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\xi}, \tag{3.5}$$

onde $\mathbf{C}$ é uma matriz $q \times p$, com $q \leq p$, de característica completa q e $\boldsymbol{\xi}$ é um vector de dimensão q previamente especificado.

Seja o caso particular:

$$H_0 : \mathbf{C}\beta_j = 0 \quad versus \quad H_1 : \mathbf{C}\beta_j \neq 0, \tag{3.6}$$

para algum j, sendo $q = 1$ e $\mathbf{C} = (0, ..., 0, 1, 0, ..., 0)$, um vector com todas as componentes nulas excepto a j-ésima que será igual a 1 e $\xi = 0$

No caso em que uma variável é policotómica e que toma $r + 1$ valores distintos, é aconselhável construir r variáveis dicotómicas para as representar havendo, nesse caso, r parâmetros $\beta's$ que lhe estão associados. Então, para averiguar se essa variável deve ou não ser incluída no modelo, interessa testar se os r parâmetros são significativamente diferentes de zero.

➔Teste de Wald

A *Estatística de Wald* (ver [Turkman e Silva, 2000]) baseia-se na normalidade assimptótica do estimador de máxima verosimilhança, $\widehat{\boldsymbol{\beta}}$.

Supõe-se que hipótese nula estabelece que $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi}$, onde $\mathbf{C}$ é uma matriz $q \times p$, com $q \leq p$, de característica completa q. Seja $\widehat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$, o qual tem uma distribuição assimptótica $N_p(\boldsymbol{\beta}, \Im^{-1}(\widehat{\boldsymbol{\beta}}))$ (aqui o vector $\boldsymbol{\beta}$ já foi substituído pela sua estimativa.), onde $\Im^{-1}(\boldsymbol{\beta})$ é matriz de covariâncias. Dado que o vector $\mathbf{C}\widehat{\boldsymbol{\beta}}$ é uma transformação linear de $\widehat{\boldsymbol{\beta}}$ então, pelas propriedades da distribuição normal multivariada,

$$\mathbf{C}\widehat{\boldsymbol{\beta}} \overset{a}{\sim} N_q\left(\mathbf{C}\boldsymbol{\beta}, \mathbf{C}\Im^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{C}^T\right)$$

e portanto, sob a hipótese nula, a estatística

$$W = \left(\mathbf{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}\right)^T \left[\mathbf{C}\Im^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{C}^T\right]^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}\right) \tag{3.7}$$

tem uma distribuição assimptótica de um χ^2 com q graus de liberdade.

A estatística W em (3.7) designa-se por *Estatística de Wald.*

Para o Teste de hipóteses referido em (3.6), designando por σ_{ii} o j-ésimo elemento da diagonal de $\Im^{-1}(\widehat{\boldsymbol{\beta}})$, a *Estatística de Wald* resume-se a:

$$W = (\widehat{\beta}_j - \beta_j)^T[\sigma_{ii}](\widehat{\beta}_j - \beta_j)$$

logo , sob H_0,

$$W = \frac{\widehat{\beta}_j^2}{\sigma_{ii}} \overset{a}{\sim} \chi_1^2.$$

Assim, rejeita-se a hipótese nula, a um nível de significância α, se o valor observado da *Estatística de Wald* for superior ao quantil de probabilidade $1-\alpha$ de um χ_q^2.

Em geral, a *Estatística de Wald* é a que se utiliza com mais frequência para testar hipóteses sobre componentes individuais, embora também se use para testar hipóteses nulas do tipo $\boldsymbol{\beta}_r = 0$ quando o subvector $\boldsymbol{\beta}_r$ representa o vector correspondente a uma recodificação de uma variável policotómica. Esta estatística é muito útil na comparação de modelos quando se começa por formar o modelo maximal e depois se consideram modelos alternativos pela exclusão de covariáveis devido, essencialmente, à utilização da estimativa não restrita de máxima verosimilhança.

➔ Teste de Razão Verosimilhanças

A *Estatística de Razão de Verosimilhanças*, também conhecida por *estatística de Wilks*, (veja-se [Turkman e Silva, 2000]), é definida por:

$$\begin{aligned} \Lambda &= -2ln\frac{max_{H_0}L(\boldsymbol{\beta})}{max_{H_0\cup H_1}L(\boldsymbol{\beta})} \\ &= -2\left\{\ell(\widetilde{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}})\right\}, \end{aligned} \tag{3.8}$$

onde $\widetilde{\boldsymbol{\beta}}$, o estimador de máxima verosimilhança restrito, é o valor de $\boldsymbol{\beta}$ que maximiza a verosimilhança sujeito às restrições impostas pela hipótese $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi}$.

O Teorema de Wilks estabelece que, sob certas condições de regularidade , a estatística Λ tem, sob H_0, uma distribuição assimptótica de um χ^2 onde o número de graus de liberdade é igual à diferença entre o número de parâmetros a estimar sob $H_0 \cup H_1$ (neste caso p) e o número de parâmetros a estimar sob H_0 (neste caso $p-q$).

Assim, sob H_0,

$$\Lambda = -2\left\{\ell(\widetilde{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}})\right\} \overset{a}{\sim} \chi^2_q$$

Com base no Teste de Razão de Verosimilhanças rejeita-se a hipótese nula H_0 : $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi}$, a um nível de significância α, se o valor observado da estatística Λ for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

A *Estatística de Razão de Verosimilhanças*, como referem [Turkman e Silva, 2000], é mais utilizada quando é preciso comparar modelos que estão encaixados, isto é, modelos em que um é submodelo do outro.

Qualidade de Ajustamento

➔ **Função Desvio**
Segundo [Turkman e Silva, 2000], o modelo completo ou saturado (S) é útil para avaliar a qualidade de ajustamento de um determinado modelo em investigação (M), através da introdução de uma medida de distância dos valores ajustados $\widehat{\mu}$ com esse modelo e dos correspondentes valores observados y. Esta medida de discrepância entre o modelo completo e o modelo corrente (modelo com menos parâmetros que está a ser sujeito a investigação) baseia-se na estatística de razão de verosimilhanças de Wilks.

O logaritmo da função de verosimilhança (*função log-verosimilhança*) de um modelo linear generalizado é dado por:

$$lnL(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{w_i\Big[y_i q(\mu_i) - b\big(q(\mu_i)\big)\Big]}{\phi} + c(y_i, \phi, w_i) \tag{3.9}$$

onde $q(\mu_i) = \theta_i$ é função de ligação canónica, w_i é um peso conhecido associado à i-ésima observação.

Comparando o modelo em investigação M com o modelo saturado S através da estatística de razão de verosimilhanças obtém-se o *desvio reduzido*:

$$\begin{aligned} D^*(\mathbf{y};\widehat{\boldsymbol{\mu}}) &= -2\Big[\ell_M(\widehat{\boldsymbol{\beta}}_M) - \ell_S(\widehat{\boldsymbol{\beta}}_S)\Big] \\ &= -2\sum_i \frac{w_i}{\phi}\bigg(\Big[y_i q(\widehat{\mu}_i) - b\big(q(\widehat{\mu}_i)\big)\Big] - \Big[y_i q(y_i) - b\big(q(y_i)\big)\Big]\bigg) \\ &= \frac{D(\mathbf{y};\widehat{\boldsymbol{\mu}})}{\phi} \end{aligned} \tag{3.10}$$

e a $D(\mathbf{y};\widehat{\boldsymbol{\mu}})$ dá-se o nome de *Desvio* para o modelo em estudo. O *Desvio* é apenas função dos dados.

É possível definir a função desvio como

$$\begin{aligned} D(\mathbf{y};\widehat{\boldsymbol{\mu}}) &= \sum_i 2w_i\Big\{y_i\Big(q(y_i)-q(\widehat{\mu}_i)\Big)-b\Big(q(y_i)\Big)+b\Big(q(\widehat{\mu}_i)\Big)\Big\} \\ &= \sum_i d_i \end{aligned} \tag{3.11}$$

a soma de parcelas d_i que medem a diferença dos logaritmos das verosimilhanças observada e ajustada para cada observação. A soma destas componentes é uma medida de discrepância total entre as duas log-verosimilhanças.

Verifica-se facilmente que o Desvio é sempre maior ou igual a zero, sendo que, para o modelo completo, é zero e vai crescendo à medida que as covariáveis vão sendo retiradas.

Outra propriedade importante do Desvio é a aditividade para modelos encaixados. Sejam dois modelos intermédios M_1 e M_2, com M_2 está encaixado em M_1, ou seja, são modelos do mesmo tipo mas o modelo M_2 contém menos parâmetros que o modelo M_1. Designando $D(\mathbf{y};\widehat{\boldsymbol{\mu}}_j)$ o desvio do modelo $M_j, j=1,2$, então a estatística da razão de verosimilhanças para comparar estes dois modelos resume-se a:

$$-2\Big(\ell_{M_2}(\boldsymbol{\beta}_2)-\ell_{M_1}(\boldsymbol{\beta}_1)\Big)=\frac{D(\mathbf{y};\widehat{\boldsymbol{\mu}}_2)-D(\mathbf{y};\widehat{\boldsymbol{\mu}}_1)}{\phi}.$$

Sob a hipótese do modelo M_1 ser verdadeiro, tem-se

$$\frac{D(\mathbf{y};\widehat{\boldsymbol{\mu}}_2)-D(\mathbf{y};\widehat{\boldsymbol{\mu}}_1)}{\phi}\overset{a}{\sim}\chi^2_{p_1-p_2}$$

onde p_j representa a dimensão do vector $\boldsymbol{\beta}$ para modelo M_j. A comparação de modelos encaixados, pode, assim, ser feita com base na diferença dos *Desvios* de cada modelo.

➔ **Coeficiente de Determinação Ajustado**

Suponha-se que se tem um modelo para o qual se pretende explicar o comportamento de uma variável Y em função de certas variáveis explicativas. Segundo [Murteira et al., 2002], estimados os parâmetros com base num determinado método obtêm-se os valores ajustados, $\hat{y}_i$, das observações da variável dependente, y_i.

Uma forma de avaliar a adequabilidade do modelo de regressão linear multipla estimado aos dados consiste em dispor de um indicador que permita medir o "grau de ajustamento" entre os y_i e os $\hat{y}_i$ $(i=1,2,\ldots,n)$. O indicador habitualmente proposto é o coeficiente de correlação entre observações da variável dependente, y_i, e os respectivos valores ajustados, $\hat{y}_i$.

O coeficiente de determinação é o quadrado do coeficiente de correlação entre os y_i e os $\hat{y}_i$ $(i=1,2,\ldots,n)$,

$$r_{y\hat{y}}^2 = \frac{\left[\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right]^2}{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}, \tag{3.12}$$

onde $\bar{y}$ e $\bar{\hat{y}}$ são as médias dos y_i e dos $\hat{y}_i$ respectivamente.

Como $0 \leq r_{y\hat{y}}^2 \leq 1$, pode concluir-se que quanto mais próximo de 1 estiver o coeficiente de determinação melhor é o "grau de ajustamento", ou seja, maior é a "proximidade" entre os y_i e os $\hat{y}_i$.

A equação (3.12) pode ser simplificada (ver [Murteira et al., 2002]), obtendo-se:

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}, \tag{3.13}$$

onde,

- SQR é a Soma dos Quadrados dos Resíduos,
- SQT é a Soma dos Quadrados Total,
- SQE é a Soma dos Quadrados Explicada.

Ao adicionar ao modelo uma variável independente, qualquer que ela seja, o coeficiente de determinação R^2 nunca decresce. Deste modo, quando se passa de k regressores para $k+1$ (acrescentando um regressor ao modelo), a introdução dessa variável adicional implica necessariamente que R^2 nunca diminua, pois SQR nunca pode crescer. Ou seja,

$$SQE \uparrow \Rightarrow R^2 \uparrow$$

Por isso, considera-se um coeficiente de determinação alternativo que desconta o efeito de um elevado número de variáveis explicativas. Este coeficiente, $\bar{R}^2$, designa-se por **Coeficiente de Determinação Ajustado** ou corrigido pelos graus de liberdade.

$$\bar{R}^2 = 1 - \frac{SQR/(T-n-1)}{SQT/(T-1)}, \tag{3.14}$$

onde T é o tamanho da amostra e n é o nº de graus de liberdade.

3.2.3 Selecção de Modelos

Segundo [Araújo, 2006], a metodologia utilizada inicialmente na construção de modelos de *Credit Scoring* era baseada essencialmente em julgamentos feitos pelos avalistas de crédito. Nestes modelos, as variáveis que compõem a classificação dos clientes e os pesos das mesmas são determinados pelos gestores de crédito da instituição bancária. Embora ainda se utilizem estes modelos em algumas instituições financeiras, actualmente a maioria dos modelos utilizados são construídos com base em técnicas estatísticas como a análise estatística multivariada, a analise discriminante e a regressão logística ou a partir de modelos de inteligência artificial, como as redes neuronais[1].

A técnica de análise discriminante foi a primeira técnica estatística utilizada na criação de modelos de *Behavioural Scoring*, no entanto, a indústria de crédito não considerou de forma séria os modelos de pontuação de crédito até meados da década de sessenta.

Só a partir da década de oitenta é que foi introduzida a técnica de regressão logística, e mais recentemente foram implantados métodos de análise de crédito baseados em redes neuronais.

Regressão Logística

A regressão Logística ou análise Logit é uma técnica estatística utilizada para produzir, a partir de um conjunto de observações, um modelo que permita a predição dos valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias.

A regressão logística é frequentemente utilizada em ciências médicas e sociais, e tem outras designações, como modelo logístico, modelo logit, e classificador de máxima entropia.

O êxito da regressão logística assenta sobretudo nas numerosas ferramentas que permitem interpretar, de modo aprofundado, os resultados obtidos. Comparando com as técnicas conhecidas em regressão, em especial a regressão linear, a regressão logística distingue-se essencialmente pelo facto de a variável resposta ser categórica.

Trata-se de um modelo de regressão para variáveis dependentes (ou de resposta) binomialmente distribuídas, $Y_i \sim B(1, \pi_i)$. É útil para modelar a probabilidade de um evento ocorrer, como função de outros factores. É um modelo linear generalizado que usa como função de ligação a função logit:

$$\theta = ln\left(\frac{\pi}{1-\pi}\right).$$

Conforme ressalta [Hair et al., 2009] a regressão logística é uma técnica estatística utilizada aquando da existência de dois grupos, e visa obter a probabilidade de uma observação pertencer a um determinado conjunto, em função do comportamento das variáveis independentes. É frequentemente utilizada para análise de dados com resposta binária ou dicotómica e consiste em relacionar, através de um modelo, a variável

[1]As redes neuronais tentam construir representações internas de modelos ou padrões detectados nos dados, os quais geralmente não são visíveis sem uma análise cuidada. As redes neuronais utilizam um conjunto de elementos de processamento (nós) análogos aos neurónios. Esses elementos de processamento são interligados numa rede que pode identificar padrões nos dados, ou seja, a rede aprende através da experiência. Em geral, em estudos de risco de crédito, utilizam-se as redes neuronais *Multi player Percepctron* através do algoritmo *Backpropagation*.

resposta (variável dependente binária) com factores que influenciam ou não a probabilidade de ocorrência de determinado evento (variáveis independentes).

Assim, na regressão logística, a variável dependente, uma vez que possui carácter não-métrico, é inserida através do uso de variáveis *dummy* (dicotómica ou binária), que assumem valor 0 para indicar a ausência de um atributo e 1 para indicar a presença de um atributo. A partir desse valor dicotómico, a regressão logística estima a probabilidade desse evento ocorrer ou não.

No âmbito da aplicação em risco de crédito, a técnica de regressão logística é utilizada para a avaliação de determinado grupo de clientes em relação ao cumprimento das suas obrigações enquanto devedor de crédito. No entanto, esta técnica assume que a probabilidade de *default* é logisticamente distribuída, com um resultado binomial 0 ou 1.

De acordo com [Hair et al., 2009], a aplicação da regressão logística implica conhecer a ocorrência ou não de determinado evento, como por exemplo, a situação de incumprimento ou não de um cliente ou a situação de insolvência ou não de uma empresa.

Podemos supor, ver [Turkman e Silva, 2000], que temos n variáveis resposta independentes $Y_i \sim B(1, \pi_i)$ ou $Y_i \sim Ber(\pi_i)$ ou seja,

$$f(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, y_i = 0, 1$$

e que a cada indivíduo i está associado um vector de especificação $\mathbf{z}_i$, que resulta do vector de covariáveis $\mathbf{x}_i$, $\quad i = 1, \ldots, n$. A distribuição Binomial pertence à Família Exponencial donde temos que $E[Y_i] = \frac{e_i^{\theta}}{1+e_i^{\theta}}$.

Como $E[Y_i] = \pi_i$ para o modelo binomial, então temos

$$\theta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right).$$

Ao fazer $\theta_i = \eta_i = \mathbf{z}_i^T\boldsymbol{\beta}$ conclui-se que a associação entre o valor esperado da variável resposta e as covariáveis é feita através da função de ligação canónica, função *logit*. Assim a probabilidade de sucesso, $\pi_i = P[Y_i = 1|X = x_i]$, está relacionada com o vector $\mathbf{z}_i$ através de

$$\pi_i = \frac{exp(\mathbf{z}_i^T\boldsymbol{\beta})}{1 + exp(\mathbf{z}_i^T\boldsymbol{\beta})} \tag{3.15}$$

Assim sendo, $Logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = ln(e_i^{\theta}) = \theta_i$ e

$$Logit(\pi_i) = \beta_0 + \beta_1 X_1 + \ldots + \beta_i X_i. \tag{3.16}$$

Como os valores possíveis de π_i se situam no intervalo $[0, 1]$, o valor de $\mathbf{z}_i^T\boldsymbol{\beta}$ é frequentemente interpretado como a probabilidade de *default*. A capacidade de estimar as probabilidades individuais constitui a principal vantagem da regressão logística. A capacidade de previsão dos modelos que utilizam as técnicas de regressão logística e os modelos de análise discriminante é similar, logo nenhuma das técnicas estatísticas é predominante sobre a outra. Na regressão Logística não é necessário verificar as suposições básicas da análise discriminante, tais como a normalidade das variáveis independentes e a igualdade das matrizes de variância/covariância para os grupos (ausência de heterocedasticidade).

Segundo [Hair et al., 2009], dever-se-ia também analisar a linearidade das relações entre as variáveis e a ausência de multicolinearidade entre as mesmas, no entanto, no presente trabalho nenhuma destas suposições serão verificadas ao pormenor, principalmente devido ao facto de a carteira ter sido obtida através da simulação, o que poderá introduzir relações entre as variáveis pouco realistas ou omitir possíveis relações que possam existir numa carteira real.

O método estatístico utilizado para a selecção das variáveis mais relevantes é o *stepwise* que verifica a ausência de multicolinearidade e corrige, automaticamente, as possíveis distorções causadas pela multicolinearidade.

Ao incluir ou excluir as variáveis independentes no modelo, uma de cada vez, com base no seu poder discriminatório, o método *stepwise* exclui as variáveis que estejam altamente correlacionadas com a variável que está a ser inserida na função.

Finalmente, é importante ressaltar que todos essas suposições são necessárias apenas para a técnica de análise discriminante. A regressão logística, por ser uma técnica mais robusta, não exige a verificação destas suposições, especialmente a normalidade.

Assim, na análise *Logit*, quando não são tidos em conta os pressupostos não implica que haja distorções dos parâmetros.

Neste trabalho apenas se utiliza a regressão Logística ou *Logit* para desenvolver o modelo de risco de crédito, *Credit Scoring*, e por esta razão, optou-se por omitir a descrição pormenorizada da análise discriminante e das redes neuronais. Vale a pena ressaltar que esta técnica é utilizada em estudos que tenham como objectivo separar dois grupos distintos, tendo como base um conjunto de variáveis independentes prédeterminadas.

Com o intuito de identificar as variáveis explicativas mais importantes foram utilizados os métodos estatísticos *Stepwise* e *AIC*, os quais serão detalhados em seguida.

Método de escolha das variáveis explicativas - Stepwise

As variáveis escolhidas inicialmente para a construção da carteira de crédito são apenas potenciais variáveis do modelo final, pois na prática, nem todas fornecem informação relevante para explicar a Qualidade de Crédito. Tornou-se assim, necessário escolher dentro desse grupo de variáveis, as mais significativas para explicar a Qualidade de Crédito, sendo que essas compõem o modelo final. Um dos métodos utilizados para realizar esta tarefa foi o método *stepwise*.

Segundo [Araújo, 2006], o método *stepwise* é utilizado, principalmente, quando se quer considerar, ao início, um número relativamente elevado de variáveis independentes para incluir na função, embora seja utilizado frequentemente em modelos iniciais nulos. A selecção da variável mais significativa na explicação da variável dependente é sequencial. Em cada passo as variáveis menos úteis, na descriminação entre os grupos, são eliminadas, ou no caso de iniciar com o modelo nulo, as variáveis mais significativas são adicionadas, e apenas são retidas um número reduzido de variáveis independentes. Esse conjunto reduzido, geralmente, é tão bom (ou melhor) que o conjunto completo de variáveis, conforme [Hair et al., 2009]. [Vasconcellos, 2002] acrescenta que a estimação *stepwise* é frequentemente utilizada em situações nas quais as variáveis independentes importantes não são conhecidas, bem como não são compreendidas as respectivas relações com a variável resposta ou dependente.

Os métodos *stepwise* podem ter duas dinâmicas diferentes a *forward stepwise* e *backward stepwise*. Sucintamente, o método *backward stepwise* parte de um modelo inicial com todas as possíveis variáveis, que vão sendo eliminadas a cada passo, até chegar ao modelo final. O método *forward stepwise* inicia-se com um modelo sem nenhuma variável explicativa e a cada passo são incluídas as variáveis relevantes até a obtenção do modelo final.

O método de selecção *stepwise*, ver [Turkman e Silva, 2000], baseia-se no valor dos *p-values* relativos aos testes de razão de verosimilhanças de Wilks entre modelos com inclusão ou exclusão de covariáveis para decidir quais as covariáveis que devem ser incluídas no modelo final. Começa por calcular o valor do *p-value* dado pelo teste de Wald e, com base neste, escolhe qual a variável que, em primeira análise, deve sair (ou entrar) no modelo final. Quanto menor (ou maior) for o valor do *p-value* dado pelo teste de Wald mais (menos) importante é considerada a covariável. Após a escolha da covariável, faz-se uma segunda análise ao seu grau de importância através do valor do *p-value* do teste de razão de verosimilhanças entre os modelos que a incluem e a excluem, e assim toma a decisão final sobre a exclusão (ou inclusão) da variável no modelo final.

Optou-se, neste trabalho, pela utilização do método *backward stepwise*.

➔Backward Stepwise

No âmbito do processo de selecção do modelo *backward*, as variáveis são removidas sequencialmente do modelo completo (que contém todas as covariáveis do estudo).

Inicialmente ajusta-se o modelo contendo todas as covariáveis e escolhe-se, com base no teste de Wald, a covariável menos significativa, isto é, a que tem o maior *p-value*. Depois, através do teste de razão de verosimilhanças, compara-se o ajuste do modelo logístico completo com o modelo que dele resulta pela exclusão da covariável menos significativa. Se o *p-value* resultante do teste de Wilks, for inferior a 0.05 volta a colocar-se a covariável no modelo, pois o modelo sem a covariável é menos explicativo do que quando a contém. Caso contrário ela é removida. Nas etapas seguintes, considera-se o modelo resultante da etapa anterior como modelo inicial e aplicam-se os mesmos passos. Caso na etapa anterior a variável não tenha sido removida e ainda haja variáveis para remover, retira-se a covariável que tem o maior *p-value* a seguir à covariável que permaneceu no modelo. Quando numa etapa todas as variáveis, pelo teste de Wald, forem significativas, sem termos em conta as que ainda assim permaneceram no modelo, o processo termina e o modelo final é constituído por todas as covariáveis dessa etapa.

Método de escolha das variáveis explicativas - AIC

O Critério de informação de Akaike foi desenvolvido em 1971 por Hirotsugu Akaike sob o nome *"Akaike Information Criterion"* (AIC) e foi proposto por Akaike em 1974. É uma medida do melhor ajuste, para um modelo estatístico estimado. É fundamentado no conceito de entropia, e oferece uma medida relativa da informação perdida quando um determinado modelo é utilizado. Pode ser usado para descrever o equilíbrio entre a variância e a tendenciosidade (*bias* ou viés) da construção do modelo, ou por outras palavras a precisão e a complexidade do modelo, ver [Wikipédia, 2009a].

O AIC não é um teste ao modelo no sentido de testar hipóteses, mas sim um teste entre modelos, ou seja, é uma ferramenta para seleccionar um modelo de entre um conjunto de modelos. Dado um conjunto de dados, e vários modelos para os mesmos, o *AIC* classifica-os e o que tiver o menor *AIC* deve ser considerado o melhor modelo.

Este critério de selecção baseia-se na função *Log-verosimilhança*, com a introdução de um factor de correcção como modo de penalização da complexidade do modelo.

A estatística correspondente para o modelo em H_0 é (ver [Turkman e Silva, 2000]),

$$AIC = -2\ell(\widetilde{\boldsymbol{\beta}}_1, 0, \widetilde{\phi}) + 2r,$$

onde $r = dim(\boldsymbol{\beta}_1)$.

A expressão (3.17) representa a relação existente entre o AIC e o *desvio reduzido* relativo ao modelo especificado por H_0 (supõe-se que o parâmetro ϕ é conhecido ou é substituído por uma estimativa consistente, neste caso de estudo em particular, será considerado $\phi = 1$).

$$\begin{aligned} AIC_r &= -2\ell(\widetilde{\boldsymbol{\beta}}_1, 0) + 2\ell(\widehat{\boldsymbol{\beta}}_S) - 2\ell(\widehat{\boldsymbol{\beta}}_S) + 2r \\ &= D_r^* + 2r - 2(\ell(\widehat{\boldsymbol{\beta}}_S)), \end{aligned} \tag{3.17}$$

onde r é o índice que específica o modelo que se está a considerar e S refere-se ao modelo completo.

É ainda sugerida uma modificação no modelo de Akaike para seleccionar modelos. Seja:

$$C_r^* = D_r^* + 2r - n = AIC_r + 2\ell(\widehat{\boldsymbol{\beta}}_S) - n.$$

Se se desenhar o gráfico de C_r^* contra r obtém-se uma boa indicação para comparação de modelos. Se o modelo for verdadeiro espera-se que C_r^* seja próximo de r.

No caso em que se tem dois modelos encaixados M_1 e M_2 com, r_1 e r_2 os respectivos parâmetros e $r_1 > r_2$ tem-se:

$$AIC_{r_1} - AIC_{r_2} = C^*_{r_1} - C^*_{r_2} = D^*_{r_1} - D^*_{r_2} + 2(r_1 - r_2),$$

e supondo que o modelo M_2 é verdadeiro, tem-se

$$\mathbb{E}[AIC_{r1} - AIC_{r2}] = r_1 - r_2 + 0(n^{-1}).$$

Na comparação de modelos sucessivamente mais ricos, o declive esperado do segmento de recta que une $AIC_{r1} - AIC_{r2}$ deve estar próximo de 1 supondo o modelo menor M_2 verdadeiro. Pares de modelos que exibem declive maior do que 1 indicam que o modelo maior não é significativamente melhor que o modelo menor.

Conforme [Wang e Salous, 2009], aumentar o número de parâmetros livres a serem estimados melhora a qualidade do ajuste, independentemente do número de parâmetros livres no processo de geração de dados.

Verifica-se então que o AIC não só mede a melhoria do modelo, como também inclui uma penalização, que é uma função crescente do número de parâmetros estimados. Esta penalidade diminui a ocorrência de *overfitting*. O modelo preferido é aquele com o menor valor de AIC. A metodologia AIC tem como objectivo encontrar o modelo que melhor explica os dados com um mínimo de parâmetros livres.

O AIC classifica um modelo pela proximidade dos seus valores com os verdadeiros valores, em termos de um determinado valor esperado. Mas é importante perceber que o valor de AIC atribuído a um modelo serve apenas para classificar os modelos concorrentes e dizer qual é o melhor entre as alternativas dadas.

3.3 Aplicação e Resultados

Nesta secção, recorrendo ao *software R*, aplicam-se os métodos estatísticos estudados para a modelação da carteira de crédito e os respectivos resultados.

Será aplicada a regressão logística para ajustar os modelos e os métodos *Stepwise-backward* e AIC para a escolha das variáveis que melhor explicam a variável resposta e que constituem o modelo final.

3.3.1 Ajustamento dos Dados Através da Regressão Logística

É importante relembrar que um modelo linear tem a forma:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + \varepsilon, \tag{3.18}$$

sendo β_i o coeficiente discriminante para a variável x_i e ε um vector de erros aleatórios.

Considera-se, inicialmente, o seguinte conjunto de variáveis: Idade, Nº Dependentes, Tipo de Habitação, Idade da Casa, Estado Civil, Código Postal, Tipo de Profissão, Antiguidade Profissão, Rendimento Líquido, Crédito, Percentagem de Entrada, Nº Total Prestações e Nº de Prestações Pagas.

Utilizando o *software R* consegue-se realizar todo o estudo que se segue.

Primeiro carregamos o conjunto de dados:

```
> data <- read.table("DADOSCR.txt", header = TRUE)
> attach(data)
```

O modelo completo será o modelo que contiver as covariáveis: *Idade, Dep, Hab, Idcasa, Civil, Postal, Prof, Antg, Rend, Cred, Entr, PTotal e Pagas*

A covariável *Código Postal* inicialmente, tal como referido na Secção 3.1.3 é uma variável categórica dividida em 4 categorias distintas: "Zona 1", "Zona 2", "Zona 3" e "Zona 4".

No entanto, após o ajustamento do modelo e a análise da significância das variáveis e respectivas categorias, o teste de Wald sugere que a categorias "Zona 2" não é significativamente diferente da categoria "Zona 1", pelo que se deverão, de forma a optimizar o modelo, agrupar estas duas categorias da variável *Código Postal.*

Os resultados do modelo inicial com quatro níveis na variável *Código Postal* poderão ser vistos nos outputs do *software R* que se encontram no Anexo B.

No que se segue, e para não alongar a exposição dos resultados, analisa-se o modelo a partir de uma variável *Código Postal* com apenas 3 categorias: "Zona 1", "Zona 3" e "Zona 4".

➔Modelo Completo

No *software R*, a função `glm` é utilizada para ajustar modelos lineares generalizados, onde se especifica quais as variáveis preditoras do modelo e a respectiva distribuição dos erros. Utilizar-se-á como função de ligação a função *Logit* (`link = "logit"`), que se omitirá nos ajustamentos seguintes pois é a função utilizada, por defeito, pelo *software R* para o modelo com erros binomiais.

```
> Modfull


Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.91e+00  -7.29e-03  -4.02e-04  -1.40e-05   2.79e+00

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.39e+01   6.07e+02   -0.07    0.942
IdadeIdosos -2.07e+00   8.73e-01   -2.37    0.018 *
IdadeJovens  3.92e+00   3.18e-01   12.31  < 2e-16 ***
DepMuitos    1.30e+00   6.88e-01    1.89    0.058 .
DepPoucos    2.84e+00   3.53e-01    8.06  7.9e-16 ***
HabComercio  1.26e-02   3.68e-01    0.03    0.973
HabEmpresa  -3.23e-01   5.99e-01   -0.54    0.590
IdcasaUsada -6.92e-02   2.60e-01   -0.27    0.790
```

```
IdcasaVelha  4.79e-01   4.90e-01    0.98    0.328
CivilDiv     3.38e+00   6.00e-01    5.63  1.8e-08 ***
CivilSol     1.60e+00   2.28e-01    7.02  2.3e-12 ***
PostalZona3  3.58e+00   6.57e-01    5.45  5.1e-08 ***
PostalZona4  5.78e+00   6.65e-01    8.69  < 2e-16 ***
ProfTipo2   -3.72e-01   2.96e-01   -1.26    0.209
ProfTipo3   -5.49e-01   2.84e-01   -1.94    0.053 .
ProfTipo4   -6.62e-01   3.62e-01   -1.83    0.068 .
AntgRecente -5.21e-01   3.81e-01   -1.37    0.172
RendBaixo    1.72e+01   6.07e+02    0.03    0.977
RendMedio    8.79e+00   6.07e+02    0.01    0.988
RendMinimo   2.37e+01   6.07e+02    0.04    0.969
Cred        -1.05e-05   1.66e-06   -6.34  2.3e-10 ***
EntrMenos10  5.11e+00   3.73e-01   13.71  < 2e-16 ***
Ptotal       3.65e-02   2.27e-03   16.11  < 2e-16 ***
Pagas       -5.67e-02   3.62e-03  -15.67  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance:  646.9  on 9976  degrees of freedom
AIC: 694.9

Number of Fisher Scoring iterations: 19
```

Realizou-se a mesma análise mas para o modelo nulo (não contém nenhuma covariável) e ainda, através do *Teste Razão de Verosimilhanças*, comparou-se qual dos dois modelos seria o que melhor explicava a variável resposta.

➔Modelo Nulo

```
> Modnull


Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.258  -0.258  -0.258  -0.258   2.615

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3871     0.0562   -60.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance: 2880.1  on 9999  degrees of freedom
AIC: 2882

Number of Fisher Scoring iterations: 6
```

➔Comparação entre os dois modelos

Considere-se o modelo `Modnull` encaixado no modelo `Modfull`. No *software R*, a função `anova`, com indicação de teste `test = "Chisq"` realiza o *Teste Razão de Verosimilhanças*, testando se o modelo com menos covariáveis é significativamente melhor que o modelo com mais covariáveis. Na prática, está-se a analisar:

H_0: Modelo `Modfull` é melhor que o modelo `Modnull`
versus
H_1: Modelo `Modnull` é melhor que o modelo `Modfull`

```
> anova(Modnull, Modfull, test = "Chisq")

Analysis of Deviance Table

Model 1: Default ~ 1
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9999       2880
2      9976        647 23     2233    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observa-se que o valor do *p-value* pelo *Teste de Wilks* é muito pequeno (inferior a 2×10^{-16}) logo, pode dizer-se que se deve rejeitar, com alguma evidência, a hipótese de que o modelo nulo é melhor que o modelo completo, ou seja, rejeita-se a hipótese de que os coeficientes acrescidos sejam iguais a zero, ou seja, pelo menos um deles será significativamente diferente de zero. Além do *Teste de Wilks*, podem-se observar os valores do Critério de Informação de Akaike (AIC) para cada um dos modelos. Verifica-se que o AIC para o modelo completo `(694.9)` é bastante menor que o do modelo nulo `(2882.1)` donde, mais uma vez se rejeita a hipótese de que o modelo nulo é significativamente melhor que o modelo completo.

Selecção de covariáveis através do Método Stepwise - Backward

O modelo completo definido anteriormente é o modelo que se utiliza para inicializar o método *Stepwise - Backward*. Este método, tal como referido na Secção 3.2.3, consiste em retirar as variáveis que o *Teste de Wald* considera menos significativas.

Observando os resultados do ajustamento do modelo completo, é possível identificar algumas covariáveis que não têm os códigos de significância. Essas são as variáveis que o *Teste de Wald* considera menos significativas, em oposição àquelas que têm '***', por exemplo, que são consideradas as mais significativas.

De acordo com o método *Stepwise*, a primeira variável a ser retirada do modelo é a que tem maior *p-value*, logo opta-se por analisar o modelo após a exclusão da variável *Rendimento* (Rend). Seguidamente compara-se o modelo completo com o modelo sem a covariável *Rend* (`a_1`) recorrendo ao *Teste Razão de Verosimilhanças.*

No que se segue, interessa testar:

H_0: Modelo `Modfull` é melhor que o modelo `a_1`
versus
H_1: Modelo `a_1` é melhor que o modelo `Modfull`.

```
> a_1 <- (Modfull - Rend)
> anova(a_1, Modfull, test = "Chisq")

Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Cred + Entr + Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9979       1484
2      9976        647  3      837    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analisando o *p-value* do teste concluímos que, embora o Teste de Wald considere a covariável *Rend* pouco significativa, o modelo com esta covariável revela-se significativamente melhor. Assim, pelo *Teste de Wilks* existe evidência em rejeitar a hipótese nula, ou seja rejeitar a hipótese de que o melhor modelo é o modelo sem a covariável *Rend*.

Para prosseguir considera-se o modelo com a covariável *Rend* incluída (Modfull) e retira-se outra covariável, nomeadamente a covariável que tem menos significância a seguir à covariável *Rend*, por outras palavras, voltando a observar os p-values dos testes de Wald para a significância dos parâmetros associados às covariáveis no modelo completo, a variável menos significativa depois da variável *Rendimento* é a variável *Tipo de Habitação* (*Hab*). Então de forma análoga faz-se:

```
> a_2 <- (Modfull - Hab)


Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.908e+00  -7.320e-03  -4.004e-04  -1.401e-05   2.796e+00

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.388e+01  6.073e+02  -0.072   0.9424
IdadeIdosos -2.093e+00  8.702e-01  -2.405   0.0162 *
IdadeJovens  3.929e+00  3.180e-01  12.353  < 2e-16 ***
DepMuitos    1.318e+00  6.864e-01   1.921   0.0548 .
DepPoucos    2.843e+00  3.527e-01   8.062 7.53e-16 ***
IdcasaUsada -7.742e-02  2.591e-01  -0.299   0.7651
IdcasaVelha  4.564e-01  4.865e-01   0.938   0.3482
CivilDiv     3.380e+00  5.997e-01   5.636 1.74e-08 ***
CivilSol     1.598e+00  2.281e-01   7.009 2.41e-12 ***
PostalZona3  3.595e+00  6.545e-01   5.492 3.97e-08 ***
PostalZona4  5.794e+00  6.623e-01   8.748  < 2e-16 ***
ProfTipo2   -3.710e-01  2.958e-01  -1.254   0.2097
ProfTipo3   -5.544e-01  2.830e-01  -1.959   0.0501 .
ProfTipo4   -6.652e-01  3.621e-01  -1.837   0.0662 .
AntgRecente -5.186e-01  3.808e-01  -1.362   0.1732
RendBaixo    1.719e+01  6.073e+02   0.028   0.9774
RendMedio    8.783e+00  6.073e+02   0.014   0.9885
RendMinimo   2.372e+01  6.073e+02   0.039   0.9688
Cred        -1.051e-05  1.650e-06  -6.371 1.88e-10 ***
EntrMenos10  5.113e+00  3.734e-01  13.693  < 2e-16 ***
Ptotal       3.649e-02  2.264e-03  16.114  < 2e-16 ***
Pagas       -5.670e-02  3.614e-03 -15.686  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance:  647.2  on 9978  degrees of freedom
AIC: 691.2

Number of Fisher Scoring iterations: 19
```

Sendo o modelo `a_2` encaixado no modelo `Modfull` considera-se o teste:

H_0: Modelo `Modfull` é melhor que o modelo `a_2`

versus

H_1: Modelo `a_2` é melhor que o modelo `Modfull`.

Aplicando o *Teste de Wilks*, verifica-se que:

```
> anova(a_2, Modfull, test = "Chisq")

Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Idcasa + Civil + Postal + Prof + Antg +
    Rend + Cred + Entr + Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9978        647
2      9976        647  2    0.302      0.86
```

Neste caso, como o *p-value* pelo *Teste de Wilks* é elevado deve rejeitar-se a hipótese de que o modelo completo é melhor que o modelo `a_2` e retira-se definitivamente a covariável *Hab*.

Prossegue-se de forma análoga, mas agora o modelo com menos covariáveis (`a_2`) passa a ser o modelo aceite no passo anterior, do qual irá ser retirada a próxima covariável e assim sucessivamente, até que se obtenha um modelo em que todas as covariáveis sejam significativas. Para não alongar a análise dos modelos encaixados indicam-se, apenas os passos intermédios sem demasiado detalhe. Para uma análise mais cuidada dos testes relativos à escolha das covariáveis, pode consultar-se o Anexo B.

```
> a_3 <- (a_2 - Idcasa)
> anova(a_3, a_2, test = "Chisq")

    p-value = P(>|Chi|)=0.6048
```

Aceita-se excluir a variável *Idcasa* e prossegue-se

```
> a_4 <- (a_3 - Prof)
> anova(a_4, a_3, test = "Chisq")

    p-value = P(>|Chi|)= 0.173
```

Aceita-se excluir a variável *Prof* e continua-se.

> a_5 <- (a_4 - Antg)
> summary(a_5)

Após a exclusão das covariáveis *Hab, Idcasa, Prof, Antg*, obtiveram-se os seguintes resultados:

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.89e+00  -7.92e-03  -4.51e-04  -1.67e-05   2.79e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37e+01   6.14e+02   -0.07    0.943
IdadeIdosos -2.15e+00   8.71e-01   -2.46    0.014 *
IdadeJovens  3.85e+00   3.12e-01   12.35  < 2e-16 ***
DepMuitos    1.27e+00   6.86e-01    1.85    0.065 .
DepPoucos    2.77e+00   3.48e-01    7.97  1.6e-15 ***
CivilDiv     3.22e+00   6.00e-01    5.36  8.1e-08 ***
CivilSol     1.56e+00   2.25e-01    6.93  4.3e-12 ***
PostalZona3  3.45e+00   6.30e-01    5.48  4.3e-08 ***
PostalZona4  5.61e+00   6.36e-01    8.82  < 2e-16 ***
RendBaixo    1.71e+01   6.14e+02    0.03    0.978
RendMedio    8.76e+00   6.14e+02    0.01    0.989
RendMinimo   2.35e+01   6.14e+02    0.04    0.969
Cred        -1.02e-05   1.63e-06   -6.28  3.3e-10 ***
EntrMenos10  5.07e+00   3.69e-01   13.73  < 2e-16 ***
Ptotal       3.60e-02   2.22e-03   16.23  < 2e-16 ***
Pagas       -5.57e-02   3.52e-03  -15.81  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  654.86  on 9984  degrees of freedom
AIC: 686.9

Number of Fisher Scoring iterations: 19
```

> anova(a_5, a_4, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Civil + Postal + Antg + Rend + Cred +
    Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9984        655
2      9983        653  1     1.67      0.20
```

Observando os resultados do *software R* para o modelo`(a_5)` verifica-se que a única covariável com pouca significância é a variável *Rendimento* cuja razão para se manter no modelo final já foi justificada. Assim, considera-se o modelo `a_5` como o modelo que melhor explica a variável dependente.

Mais concretamente, o melhor modelo encontrado através do método Stepwise-Backward é constituído pelas covariáveis : *Idade, Dep, Civil, Postal, Cred, Entr, Ptotal, Pagas e Rend.* Desta forma podemos ajustar o modelo Logit ao modelo `a_5`, designando-o por `ModStp`.

> ModStp

Os valores do output do *software R* para o modelo `ModStp` são os mesmos descritos pelo output para o modelo `a_5`.

Selecção de covariáveis através do Critério de Informação de Akaike

Um outro método de selecção de modelos é o Critério de Informação de Akaike (AIC) já descrito anteriormente na Secção 3.2.3. Nesta secção recorre-se a este método para seleccionar as covariáveis que melhor explicam a variável resposta e assim obter o melhor modelo.

O *software R* contém uma função que executa todos os passos necessários para obter a melhor escolha de covariáveis, através de critério de Akaike.
Primeiro passo AIC:

```
Start:  AIC=694.9
Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas

         Df Deviance     AIC
- Hab     2   647.20  691.20
- Idcasa  2   647.96  691.96
- Prof    3   651.61  693.61
- Antg    1   648.84  694.84
<none>        646.90  694.90
- Cred    1   694.41  740.41
- Civil   2   719.16  763.16
- Dep     2   737.00  781.00
- Postal  2   846.73  890.73
- Idade   2   861.43  905.43
- Entr    1  1048.60 1094.60
- Ptotal  1  1416.60 1462.60
- Rend    3  1483.54 1525.54
- Pagas   1  1791.06 1837.06
```

Retira a covariável *Hab* e obtém:

```
Step:  AIC=691.20
Default ~ Idade + Dep + Idcasa + Civil + Postal + Prof + Antg +
    Rend + Cred + Entr + Ptotal + Pagas
```

De forma análoga vai retirar a covariável *Idcasa* donde resulta um `AIC=688.2`, depois a covariável *Prof* obtendo `AIC=687.19`, seguida da covariável *Antg* obtendo-se `AIC=686.86` donde resulta o modelo final:

```
Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
         Df Deviance     AIC
<none>        654.86  686.86
+ Antg    1   653.19  687.19
+ Prof    3   650.08  688.08
+ Idcasa  2   653.75  689.75
+ Hab     2   654.51  690.51
- Cred    1   701.22  731.22
- Civil   2   723.77  751.77
```

```
- Dep     2   742.47  770.47
- Postal  2   851.66  879.66
- Idade   2   867.88  895.88
- Entr    1  1056.59 1086.59
- Ptotal  1  1422.12 1452.12
- Rend    3  1487.42 1513.42
- Pagas   1  1798.61 1828.61
Residual Deviance: 654.9        AIC: 686.9
```

Note-se que as variáveis que o método AIC opta por retirar são as mesmas que foram retiradas pelo método Stepwise.

Designando o modelo definido pelo critério AIC por `ModAIC` e fazendo o seu ajustamento tem-se:

```
> ModAIC


Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.89e+00  -7.92e-03  -4.51e-04  -1.67e-05   2.79e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37e+01   6.14e+02   -0.07    0.943
IdadeIdosos -2.15e+00   8.71e-01   -2.46    0.014 *
IdadeJovens  3.85e+00   3.12e-01   12.35  < 2e-16 ***
DepMuitos    1.27e+00   6.86e-01    1.85    0.065 .
DepPoucos    2.77e+00   3.48e-01    7.97  1.6e-15 ***
CivilDiv     3.22e+00   6.00e-01    5.36  8.1e-08 ***
CivilSol     1.56e+00   2.25e-01    6.93  4.3e-12 ***
PostalZona3  3.45e+00   6.30e-01    5.48  4.3e-08 ***
PostalZona4  5.61e+00   6.36e-01    8.82  < 2e-16 ***
RendBaixo    1.71e+01   6.14e+02    0.03    0.978
RendMedio    8.76e+00   6.14e+02    0.01    0.989
RendMinimo   2.35e+01   6.14e+02    0.04    0.969
Cred        -1.02e-05   1.63e-06   -6.28  3.3e-10 ***
EntrMenos10  5.07e+00   3.69e-01   13.73  < 2e-16 ***
Ptotal       3.60e-02   2.22e-03   16.23  < 2e-16 ***
Pagas       -5.57e-02   3.52e-03  -15.81  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  654.86  on 9984  degrees of freedom
AIC: 686.86

Number of Fisher Scoring iterations: 19
```

Em ambos os métodos, o Teste de Wald sugere que a categoria "Muitos" da covariável *Dep* (Nº de Dependentes) não é significativamente diferente da categoria "Alguns". Optou-se, no entanto, por manter as categorias separadas porque se considerou inicialmente, na simulação da carteira de crédito, que os clientes na categoria "Jovens" da covariável *Idade* só poderiam ter "Poucos" ou "Alguns" dependentes logo agrupar a categoria "Alguns" com a categoria "Muitos" iria contra o pressuposto inicial da simulação.

```
> anova(ModAIC, Modfull, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9984        654.86
2      9976        646.90  8     7.9635      0.437
```

Observando os resultados da `anova` entre o Modelo AIC e o Modelo Completo, concluímos que o Modelo AIC (`ModAIC`) é significativamente melhor que `Modfull`.

No entanto como a covariável *Rend* não é significativa pelo Teste de Wald, vamos retirá-la e analisar a `anova` dos dois modelos. Seja `ModAIC2` o Modelo AIC sem a variável *Rend*, ou seja, o modelo `ModAIC2` encaixado no modelo `ModAIC` e realiza-se o teste:

H_0: Modelo `ModAIC` é melhor que o modelo `ModAIC2`

versus

H_1: Modelo `ModAIC2` é melhor que o modelo `ModAIC`.

```
> ModAIC2 <-(ModAIC - Rend)
> anova(ModAIC2, ModAIC, test = "Chisq")
```

```
AIC: 1513.4
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9987       1487.42
2      9984        655.86  3      832.56    < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Verifica-se novamente que, embora a variável *Rend* não seja significativa pelo Teste de Wald, o modelo com esta variável inclusa é significativamente melhor. Logo, no modelo final optou-se por incluir a variável *Rend*, até porque, de um ponto de vista prático parece-nos que a instituição bancária considerará o valor do rendimento de um cliente como um factor relevante na análise de concessão de crédito.

A razão que leva a esta diferença de decisão entre os diferentes testes deve-se, em princípio, ao facto da variável *Rend* ter sido simulada como uma variável categórica e não uma variável numérica à qual, no entanto, se atribuiu tem peso bastante elevado (ver novamente Tabela 3.3), devido às razões referidas anteriormente.

Coeficiente de Determinação Ajustado ($\bar{R}^2$)

O Coeficiente de Determinação Ajustado descrito na Secção 3.2.2, será calculado, nesta secção, para os modelos estimados anteriormente, de forma a analisar a qualidade dos ajustamentos com base noutro critério.

Coeficiente de Determinação Ajustado:

$$\bar{R}^2 = 1 - \frac{SQR/(T-n-1)}{SQT/(T-1)}. \tag{3.19}$$

A Tabela 3.6 contém os valores dos Coeficientes de Determinação Ajustados e dos Critérios de Akaike para os modelos ajustados anteriormente.

Modelo	$\bar{R}^2$	**AIC**
Modelo Completo	0.7757632	694.8967
Modelo AIC (ou Stepwise)	0.7722856	686.8602
Modelo AIC sem covariável Rendimento	0.4829361	1513.418

Tabela 3.6: Coeficientes de Determinação Ajustados e Critérios de Akaike

Observando a Tabela 3.6 conclui-se que o Coeficiente de Determinação Ajustado é melhor para o modelo completo do que para os modelos estimados pelos métodos *Stepwise* e *Critério de Akaike* (os modelos estimados por estes dois métodos contêm as mesmas variáveis explicativas), embora com uma diferença pequena comparada com o facto de se ter obtido um modelo mais simples. O valor do Critério de Informação de Akaike revela que estes últimos modelos são significativamente melhores que o modelo completo. Em relação ao modelo estimado sem a covariável *Rendimento*, observa-se uma diferença considerável, tanto para o Coeficiente de Determinação Ajustado como para o Critério de Akaike e conclui-se que este modelo é bastante pior que os restantes.

Confirmação Selecção de Variáveis

Para confirmar a selecção de variáveis pelas quais se optou, utilizaram-se duas funções do *software R*.

Através do teste χ^2, uma das funções analisa quais as variáveis do modelo que são mais significativas. Obtiveram-se as mesmas covariáveis que já haviam sido escolhidas através nos métodos anteriores conforme se pode verificar nos outputs do Anexo B.

Considerou-se ainda a possibilidade da existência de interacções entre algumas das covariáveis. Para analisar eventuais relações de 2ª ordem entre covariáveis que pudessem melhorar o modelo utilizou-se a outra função.

```
Model:
Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof + Antg + Rend + Cred + Entr
        + Ptotal + Pagas
```

	Df	Deviance	AIC	LRT	Pr(Chi)	
<none>		646.90	694.90			
Idade:Dep	3	645.68	699.68	1.2177	0.748773	
Idade:Hab	4	642.22	698.22	4.6807	0.321657	
Idade:Idcasa	4	645.08	701.08	1.8196	0.768895	
Idade:Civil	4	645.76	701.76	1.1333	0.888946	
Idade:Postal	4	643.83	699.83	3.0696	0.546244	
Idade:Prof	6	644.75	704.75	2.1456	0.905835	
Idade:Antg	2	646.49	698.49	0.4041	0.817056	
Idade:Rend	6	646.61	706.61	0.2888	0.999550	
Idade:Cred	2	640.01	692.01	6.8834	0.032010	*
Idade:Entr	2	646.31	698.31	0.5849	0.746421	
Idade:Ptotal	2	646.03	698.03	0.8626	0.649679	
Idade:Pagas	2	645.12	697.12	1.7790	0.410859	
Dep:Hab	4	637.11	693.11	9.7838	0.044231	*
Dep:Idcasa	4	644.46	700.46	2.4370	0.655951	
Dep:Civil	4	643.77	699.77	3.1289	0.536489	
Dep:Postal	4	644.51	700.51	2.3888	0.664644	
Dep:Prof	6	642.76	702.76	4.1349	0.658433	
Dep:Antg	2	646.56	698.56	0.3317	0.847158	
Dep:Rend	6	645.88	705.88	1.0175	0.984941	
Dep:Cred	2	644.35	696.35	2.5456	0.280041	
Dep:Entr	2	643.27	695.27	3.6235	0.163368	
Dep:Ptotal	2	641.53	693.53	5.3693	0.068245	.
Dep:Pagas	2	640.64	692.64	6.2595	0.043729	*
Hab:Idcasa	4	642.99	698.99	3.9041	0.419144	
Hab:Civil	4	633.10	689.10	13.7947	0.007980	**
Hab:Postal	4	642.99	698.99	3.9020	0.419430	
Hab:Prof	6	638.02	698.02	8.8773	0.180595	
Hab:Antg	2	644.92	696.92	1.9807	0.371442	
Hab:Rend	6	643.84	703.84	3.0535	0.802099	
Hab:Cred	2	646.21	698.21	0.6871	0.709266	
Hab:Entr	2	644.55	696.55	2.3441	0.309727	
Hab:Ptotal	2	645.27	697.27	1.6232	0.444147	
Hab:Pagas	2	643.24	695.24	3.6596	0.160443	
Idcasa:Civil	4	642.75	698.75	4.1432	0.386973	
Idcasa:Postal	4	643.34	699.34	3.5534	0.469802	
Idcasa:Prof	6	640.42	700.42	6.4742	0.372211	
Idcasa:Antg	2	645.97	697.97	0.9250	0.629703	
Idcasa:Rend	6	643.90	703.90	2.9955	0.809408	
Idcasa:Cred	2	646.02	698.02	0.8717	0.646703	
Idcasa:Entr	2	645.32	697.32	1.5718	0.455700	
Idcasa:Ptotal	2	646.61	698.61	0.2884	0.865715	
Idcasa:Pagas	2	646.05	698.05	0.8494	0.653957	
Civil:Postal	4	636.18	692.18	10.7169	0.029937	*
Civil:Prof	6	641.24	701.24	5.6553	0.462884	
Civil:Antg	2	646.08	698.08	0.8200	0.663646	
Civil:Rend	6	644.60	704.60	2.2945	0.890724	
Civil:Cred	2	645.50	697.50	1.4003	0.496511	
Civil:Entr	2	643.80	695.80	3.1003	0.212217	
Civil:Ptotal	2	643.82	695.82	3.0804	0.214336	
Civil:Pagas	2	645.95	697.95	0.9494	0.622070	
Postal:Prof	6	645.15	705.15	1.7460	0.941512	
Postal:Antg	2	645.03	697.03	1.8668	0.393218	
Postal:Rend	6	642.67	702.67	4.2306	0.645497	
Postal:Cred	2	646.55	698.55	0.3441	0.841925	
Postal:Entr	2	646.30	698.30	0.5934	0.743280	
Postal:Ptotal	2	646.01	698.01	0.8885	0.641320	
Postal:Pagas	2	642.66	694.66	4.2410	0.119974	
Prof:Antg	3	645.06	699.06	1.8352	0.607295	
Prof:Rend	9	644.46	710.46	2.4402	0.982449	
Prof:Cred	3	641.07	695.07	5.8291	0.120228	
Prof:Entr	3	646.04	700.04	0.8517	0.837067	
Prof:Ptotal	3	645.66	699.66	1.2321	0.745316	
Prof:Pagas	3	645.33	699.33	1.5679	0.666699	
Antg:Rend	3	645.83	699.83	1.0651	0.785501	
Antg:Cred	1	646.13	696.13	0.7686	0.380665	
Antg:Entr	1	646.48	696.48	0.4161	0.518908	
Antg:Ptotal	1	642.84	692.84	4.0582	0.043957	*
Antg:Pagas	1	642.29	692.29	4.6038	0.031901	*
Rend:Cred	3	643.61	697.61	3.2828	0.350045	
Rend:Entr	3	646.85	700.85	0.0509	0.996988	
Rend:Ptotal	3	646.88	700.88	0.0134	0.999588	
Rend:Pagas	3	645.00	699.00	1.8948	0.594533	
Cred:Entr	1	646.67	696.67	0.2248	0.635396	
Cred:Ptotal	1	641.38	691.38	5.5140	0.018865	*
Cred:Pagas	1	639.02	689.02	7.8771	0.005007	**
Entr:Ptotal	1	644.95	694.95	1.9431	0.163335	
Entr:Pagas	1	645.72	695.72	1.1763	0.278110	
Ptotal:Pagas	1	630.66	680.66	16.2409	5.578e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analisando os resultados, verifica-se que o teste efectuado indica como sendo bastante significativa, a relação entre as covariáveis *Pagas* e *Ptotal*. Isso implica que o modelo a considerar deveria ser:

$$Default \sim Idade + Dep + Civil + Postal + Rend + Cred + Entr + Ptotal \times Pagas.$$

Esta sugestão deriva da utilização da variável $Pagas/PTotal$ aquando da construção da variável resposta *Default* durante a simulação da carteira de crédito, variável à qual foi atribuído um peso elevado.

Como tal, os resultados do Teste Razão de Verosimilhanças demonstram uma melhoria na qualidade do modelo quanto se tem em conta esse factor.

No entanto, optou-se pela não inclusão desta relação entre covariáveis, uma vez que esta poderá provir da simulação bem como, de um ponto de vista práctico, esta relação não fará muito sentido.

Uma segunda sugestão, com elevado nível de significância, que resultou da aplicação desta função foi multiplicar a variável *Cred* pela variável *Pagas* o que, da mesma forma, poderá ser fruto de simulação e não faz muito sentido do ponto de vista prático. No entanto, mais uma vez, a introdução desta relação traduz-se numa melhoria na qualidade do modelo.

3.3.2 Estimação da Probabilidade de Default

Após o estudo estatístico dos dados e aplicação de métodos de selecção de covariáveis encontrou-se o modelo final (`ModStp` ou `ModAIC`) que melhor explica a variável *Default*. Este modelo contempla as variáveis *Idade, Nº de Dependentes, Estado Civil, Código Postal, Rendimento Líquido, Crédito, Percentagem de Entrada, Nº de Prestações Totais e Nº de Prestações Pagas.*

Do ajustamento do modelo final resultam os coeficientes $\beta_i's$ que serão utilizados, nesta secção, para a estimação da probabilidade de *default*. A Tabela 3.7 contém os valores dos coeficientes, $\beta_i's$, a considerar.

Característica	**Coeficientes**	
(Intercept)	β_0	-43.74
IdadeJovens	β_1	3.851
IdadeIdosos	β_2	-2.146
DepPoucos	β_3	2.769
DepMuitos	β_4	1.267
CivilDiv	β_5	3.219
CivilSol	β_6	1.556
PostalZona3	β_7	3.450
PostalZona4	β_8	5.608
RendMinimo	β_9	23.500
RendBaixo	β_{10}	17.060
RendMedio	β_{11}	8.765
Cred	β_{12}	-1.023×10^{-5}
EntrMenos10	β_{13}	5.073
Ptotal	β_{14}	3.595×10^{-2}
Pagas	β_{15}	-5.568×10^{-02}

Tabela 3.7: Coeficientes do Modelo Ajustado

Note-se que os valores dos coeficientes da Tabela 3.7, são os valores que resultam do ajustamento do modelo `ModStp` ou `ModAIC` (veja-se Secção 3.3.1). Convém ainda ressaltar que algumas categorias das variáveis contempladas, neste modelo, estão subjacentes a β_0, designadamente as categorias: *IdadeAdultos*, *DepAlguns*, *CivilCas*, *PostalZona1* (que contempla a *PostalZona2* inicial), *RendAlto* e *EntrMais10*.

Então, pela equação (3.18) tem-se:

$$\begin{aligned} Y = \beta_0 &+ \beta_1\, IdadeJovens + \beta_2\, IdadeIdosos + \beta_3\, DepPoucos + \\ &+ \beta_4\, DepMuitos + \beta_5\, CivilDiv + \beta_6\, CivilSol + \\ &+ \beta_7\, PostalZona3 + \beta_8\, PostalZona4 + \beta_9\, RendMinimo + \\ &+ \beta_{10}\, RendBaixo + \beta_{11}\, RendMedio + \beta_{12}\, Cred + \\ &+ \beta_{13}\, EntrMenos10 + \beta_{14}\, PTotal + \beta_{15}\, Pagas. \end{aligned} \tag{3.20}$$

A título de exemplo, para um cliente com as características da Tabela 3.8, o valor de Y, de acordo com a equação (3.18), obtém-se da seguinte forma:

$\widehat{Y} = \beta_0 + \beta_3\, DepPoucos + \beta_6\, CivilSol + \beta_7\, PostalZona3 + \beta_9\, RendMinimo + \beta_{12}\, Cred + \beta_{13}\, EntrMenos10 + \beta_{14}\, Ptotal + \beta_{15}\, Pagas$

Idade	**Dep**	**Civil**	**Postal**	**Rend**	**Cred**	**Entr**	**Ptotal**	**Pagas**
Adultos	Poucos	Sol	Zona3	Minimo	159684	Menos10	480	94

Tabela 3.8: Características de Um Cliente - Exemplo

Finalmente, para o cálculo da probabilidade de cada cliente, π_i, utiliza-se a equação (3.15), definida na Secção 3.2.3. A equação (3.15), pode ser reescrita da seguinte forma:

$$\pi_i = \frac{e_i^Y}{1 + e_i^Y}. \tag{3.21}$$

Na Tabela 3.9 estão as probabilidades de *default* estimadas, conforme se tem vindo a descrever, para clientes com diferentes características. Nesta Tabela consta também o valor do *spread* mínimo que se deve atribuir a cada cliente, estimado com base na probabilidade de *default* respectiva e numa taxa de recuperação constante igual a 0.85, segundo a Equação (2.27).

	Características										$\widehat{\boldsymbol{\pi}}_i$	$\widehat{\mathbf{s}}$
Cliente	**Idade**	**Dep**	**Civil**	**Postal**	**Rend**	**Cred**	**Entr**	**Ptotal**	**Pagas**	$\widehat{y}_i$	$\frac{e^{\widehat{y}_i}}{1+e^{\widehat{y}_i}}$	$\widehat{\pi}_i(1-\widehat{R})$
A	Jovens	Poucos	Sol	Zona1	Medio	109930	Mais10	240	98	-24.75	0.00%	0.00%
B	Idosos	Alguns	Sol	Zona1	Minimo	113420	Mais10	240	199	-24.44	0.00%	0.00%
C	Adultos	Alguns	Cas	Zona4	Baixo	292247	Menos10	240	158	-19.16	0.00%	0.00%
D	Jovens	Poucos	Cas	Zona1	Medio	179610	Mais10	480	73	-17.00	0.00%	0.00%
E	Jovens	Poucos	Sol	Zona4	Alto	262039	Menos10	420	63	-15.97	0.00%	0.00%
F	Adultos	Poucos	Cas	Zona3	Minimo	79537	Menos10	360	143	-4.78	0.83%	0.12%
G	Jovens	Poucos	Cas	Zona4	Minimo	107710	Mais10	300	82	-2.89	5.24%	0.79%
H	Adultos	Alguns	Sol	Zona3	Minimo	154810	Menos10	420	91	-1.71	15.28%	2.29%
I	Jovens	Poucos	Div	Zona4	Medio	178414	Menos10	540	70	-0.76	31.76%	4.76%
J	Adultos	Muitos	Sol	Zona4	Minimo	170030	Menos10	480	171	-0.74	32.29%	4.84%
K	Adultos	Poucos	Cas	Zona4	Minimo	127746	Menos10	480	161	0.19	54.85%	8.23%
L	Jovens	Poucos	Div	Zona4	Medio	89984	Menos10	540	60	0.70	66.74%	10.01%
M	Idosos	Poucos	Cas	Zona4	Minimo	88547	Menos10	420	70	1.36	79.57%	11.94%
N	Adultos	Poucos	Sol	Zona3	Minimo	159684	Menos10	480	94	3.00	95.24%	14.29%
O	Jovens	Poucos	Sol	Zona4	Minimo	211987	Menos10	360	108	3.38	96.70%	14.50%

Tabela 3.9: Probabilidades de *Default* Estimadas - Exemplos

3.3.3 Avaliação da Capacidade Preditiva do Modelo

Para validar os resultados obtidos deve-se, antes da aplicação dos modelos, dividir a carteira de crédito em duas amostras, a amostra de desenvolvimento e a amostra de validação. O objectivo é avaliar se o modelo construído a partir da amostra de desenvolvimento, classifica correctamente as observações que não foram utilizadas no processo de estimação. Desta forma, a carteira de clientes (12000 clientes) simulada foi dividida numa amostra de desenvolvimento (10000 clientes) utilizada nas secções anteriores e numa amostra de teste (2000 clientes) que é utilizada nesta secção para avaliar a capacidade preditiva do modelo `ModAIC` estimado.

A avaliação da capacidade preditiva dos modelos foi realizada com a ajuda de matrizes de classificação. Estas matrizes de classificação consistem numa comparação entre a classificação (cumpridores ou incumpridores) dos 2000 clientes da amostra de teste e a probabilidade de *default* estimada pelo modelo Logístico para os mesmos clientes.

A estimação da probabilidade de *default* dos clientes da amostra de teste fez-se de forma análoga à estimação da probabilidade de *default* dos clientes da carteira (ver Secção 3.3.2).

Nas tabelas que se seguem o *Grupo C* designa o grupo de *clientes Cumpridores* e o *Grupo I*, os *clientes Incumpridores*. As percentagens a negrito representam a percentagem de clientes bem classificados, ou seja, os clientes que após a aplicação do modelo logístico se classificaram correctamente quanto a serem cumpridores ou incumpridores.

Para a construção das matrizes de classificação consideraram-se dois casos: no primeiro, considerou-se que os clientes com uma probabilidade de *default* estimada superior a 50% eram clientes incumpridores; no segundo caso, considerou-se que os clientes com probabilidade de *default* estimada superior a 25% eram clientes incumpridores.

Desta forma, numa primeira análise, contabilizaram-se os clientes da amostra de teste que eram incumpridores (1) e os clientes cumpridores (0) e para cada um destes dois grupos, os clientes cuja probabilidade de *default* estimada era superior ou igual a 50% e inferior a 50%, como se pode observar na Tabela 3.10.

Original	**Classificação do Modelo**		
	< 50	≥ 50	
0	**1908**	**8**	1916
1	**29**	**55**	84
	1937	63	**Total**

Tabela 3.10: Contabilização de Clientes Cumpridores e Incumpridores - 50%

	Classificação do Modelo	
Classe Original	Grupo C	Grupo I
Grupo C	**99.58%**	0.42%
Grupo I	34.52%	**65.48%**

Tabela 3.11: Matriz de Classificação do Modelo de Aprovação de Crédito (50%)

Observando as Tabelas 3.10 e 3.11 pode dizer-se que o modelo utilizado para a estimação das probabilidades de *default* está a classificar correctamente cerca de 99.58% dos clientes cumpridores e 65.48% dos clientes incumpridores, se se tiver em conta que apenas se consideram clientes incumpridores os clientes com probabilidade de default estimada superior ou igual a 50%. Donde, em média, o modelo classifica correctamente 82.53% dos clientes da amostra de teste.

Numa segunda análise, considerou-se que os clientes cuja probabilidade de *default* estimada era superior ou igual a 25% eram considerados incumpridores e construíram-se as Tabelas 3.12 e 3.13.

Original	Classificação do Modelo		
	< 25	≥ 25	
0	**1891**	**25**	1916
1	**18**	**66**	84
	1909	91	Total

Tabela 3.12: Contabilização de Clientes Cumpridores e Incumpridores - 25%

	Classificação do Modelo	
Classe Original	Grupo C	Grupo I
Grupo C	**98.70%**	1.30%
Grupo I	21.43%	**78.57%**

Tabela 3.13: Matriz de Classificação do Modelo de Aprovação de Crédito (25%)

Nesta última análise, pode dizer-se que o modelo utilizado para a estimação das probabilidades de *default* está a classificar correctamente cerca de 98.70% dos clientes cumpridores e 78.57% dos clientes incumpridores, portanto, em média, o modelo classifica correctamente 88.63% dos clientes da amostra de teste.

Observa-se também, que o modelo é também mais eficiente a classificar clientes cumpridores. Este resultado deve-se, talvez, ao facto de na simulação da variável *Default*, se ter optado por apenas considerar que os clientes com uma ponderação superior a 85.5% (veja-se Expressão (3.1)), fossem clientes incumpridores, ou seja, foram excluídos clientes que, eventualmente, estão a ser classificados no modelo final como incumpridores.

Verifica-se, de uma forma geral, que o modelo desenvolvido utilizando a regressão logística (`ModAIC`) obteve bons resultados na classificação de clientes como cumpridores e incumpridores.

3.3.4 Análise de Resíduos

Após o ajustamento de um modelo de dados, deve-se investigar quão bem o modelo consegue descrever os dados. Em particular, deve-se procurar a existência de alguma tendência sistemática na qualidade do ajustamento. Por exemplo, se a qualidade do ajustamento aumenta com o aumento do número de observações, ou se é função de uma ou mais variáveis explicativas.

Segundo [Turkman e Silva, 2000], a análise de resíduos é útil, não só para realizar uma avaliação local da qualidade de ajustamento de um modelo no que diz respeito à escolha da distribuição, da função de ligação e de termos do preditor linear, como também para ajudar a identificar observações mal ajustadas, ou seja, que não são bem explicadas pelo modelo. Um resíduo R_i deve exprimir a discrepância entre o valor observado y_i e o valor $\widehat{y}_i$ ajustado pelo modelo.

Não tendo como objectivo, neste trabalho, fazer um estudo intensivo sobre a análise de resíduos serão apresentados aqui alguns gráficos resultantes da análise de resíduos do modelo estimado e breves observações sobre os mesmos. Numa carteira de crédito real seria essencial uma análise de resíduos mais aprofundada.

Convém referir que existe alguma ambiguidade em definir os resíduos na regressão logística (ver [Jennings, 1986], [Bedrick e Hill, 1990] e [Duffy, 1990]).

Na análise de resíduos, existem vários tipos de resíduos dos quais se destacam os resíduos de Pearson e os desvios residuais (ou *Deviance Residuals*).

Considere-se os dados $(\mathbf{Y}_i, m_i, x_i) : 1 \leq i \leq n$, onde o $\mathbf{Y}_i$ são independentemente distribuídas com distribuição Binomial(m_i, π_i), e as π_i satisfazem $logit(\pi_i) = log\left[\pi_i/(1-\pi_i)\right] = \mathbf{x}_i^T\boldsymbol{\beta}$, com $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ik})^T$ vectores covariável conhecidos e $\boldsymbol{\beta} \in \mathbb{R}^k$ desconhecido. Seja $\mathbf{X}$ a matriz $n \times k$ de dimensão k com x_i' a $i-$ésima linha, e $\mathbf{D}$ denota a matriz diagonal $n \times n$ com $m_i\pi_i(1-\pi_i)$ o $i-$ésimo elemento diagonal. Define-se ainda

$$\mathbf{H} = \mathbf{D}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^{\frac{1}{2}}. \tag{3.22}$$

Considere-se um resíduo $R(y_i, p_i)$, função da observação y_i e da probabilidade p_i.

A expressão para os resíduos de Pearson, segundo [Duffy, 1990], podem-se escrever da seguinte forma:

$$R_i^P(y_i, \pi_i) = \frac{(y_i - m_i\pi_i)}{\left[m_i\pi_i(1-m_i\pi_i)\right]^{1/2}}. \tag{3.23}$$

O desvio residual é a medida do desvio do contributo de cada observação e pode ser utilizado para verificar, nos modelos lineares generalizados, a adequação do modelo ajustado para cada observação e podem-se determinar da seguinte forma:

$$R_i^D(y_i, \pi_i) = \delta_i\left[2\left(y_i \ln\left(\frac{y_i}{m_i\pi_i}\right) + (m_i - y_i)\ln\left(\frac{m_i - y_i}{m_i - m_i\pi_i}\right)\right)\right]^{1/2}, \tag{3.24}$$

onde $\delta_i = sinal(y_i - m_i\pi_i)$ e se

$$\begin{cases} y_i = 0 & \Rightarrow R_i^D(y_i, \pi_i) = -\left\{2m_i|\log(1-\pi_i)|\right\}^{\frac{1}{2}} \\ y_i = m_i & \Rightarrow R_i^D(y_i, \pi_i) = \left\{2m_i|\log(\pi_i)|\right\}^{\frac{1}{2}}. \end{cases} \tag{3.25}$$

[McCullagh e Nelder, 1989] sugerem a utilização dos desvios residuais R_i^D em vez dos resíduos de Pearson, já que estes, embora apresentem propriedades de segunda ordem razoáveis, podem ter distribuições muito distintas da Normal.

Segundo [Crawley, 2007], uma das razões mais comuns para a falta de ajuste é através da existência de *outliers* nos dados. No entanto, é importante entender que um ponto pode parecer ser um *outlier* devido à má especificação do modelo, e não porque haja algo de errado com os dados.

Para analisar o gráfico da Figura 3.8 deve ter-se em conta alguns conceitos, como alavancagem, influência e distância de Cook. Tais conceitos serão brevemente explicados a seguir.

Para [Crawley, 2007], em estatística, alavancagem (*leverage*) é um termo utilizado em conexão com a análise de regressão e, em particular, em análises que visam identificar as observações que têm um grande efeito sobre o resultado da medição de modelos de regressão. Os pontos de alavancagem são as observações ou os valores das variáveis independentes tais que a falta de observações vizinhas significa que o modelo de regressão passará próximo dessa observação particular. As observações alavancadas encontram-se distanciadas do valor médio de y. Medidas de alavancagem de um determinado ponto x são proporcionais a $(y-\overline{y})^2$. A medida de alavancagem mais comum é:

$$h_i = \frac{1}{n} + \frac{(y_i-\overline{y})^2}{\sum (y_j-\overline{y})^2}, \tag{3.26}$$

com h_i o $i-$ésimo elemento da diagonal da matriz $\mathbf{H}$ definida em (3.22).

É comum dizer-se que um ponto é muito influente se:

$$h_i > \frac{2p}{n}, \tag{3.27}$$

onde p é o número de parâmetros no modelo e n o número de observações.

A influência dos dados não pode, de todo, ser analisada apenas através dos resíduos, pois um ponto muito influente pode forçar a linha de regressão a aproximar-se dele e, portanto, o ponto influente pode ter um resíduo pequeno. Observações influentes são observações que, se retiradas da análise, gerariam variações assinaláveis no conjunto dos valores ajustados de Y e dos parâmetros estimados, $\widehat{\boldsymbol{\beta}}$.

Uma medida frequente para a influência da observação i é a distância de Cook. A distância de Cook representa o efeito de excluir um dado observado e é também uma tentativa de combinar a alavancagem e os resíduos numa única medida. Os pontos com resíduos elevados (*outliers*) e/ou com alta alavancagem podem distorcer a precisão de uma regressão. A distância de Cook, em valor absoluto, é dada por:

$$C_i = |r_i^*| \left(\frac{n-p}{p}\cdot\frac{h_i}{1-h_i}\right)^{1/2}. \tag{3.28}$$

Observações atípicas, influentes ou alavancadas, embora possam estar relacionadas, não são o mesmo conceito. Por exemplo, uma observação com resíduo *standardizado* grande e h_i elevado, tem de ter uma distância de Cook grande, ou seja ser influente. Se tiver r_i grande e h_i pequeno, pode, ou não, ser influente, consoante a grandeza relativa desses dois valores. Estes diagnósticos servem sobretudo para identificar observações que merecem maior atenção e consideração.

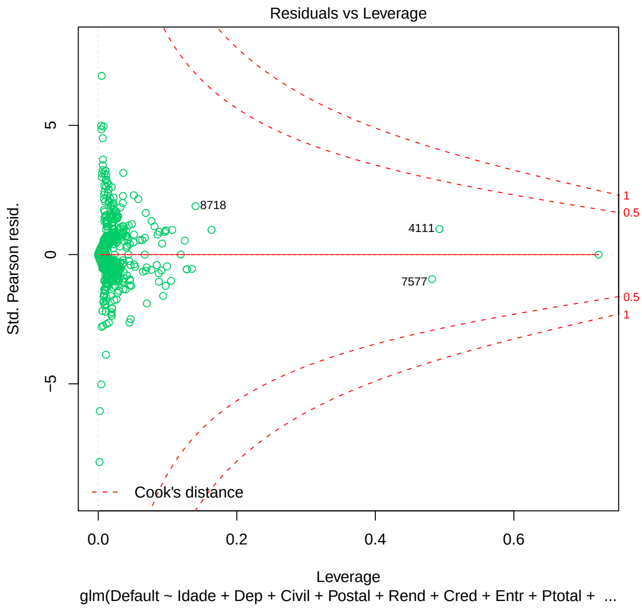

Figura 3.8: Residuos de Pearson Padronizados $\times$ Alavancagem

O gráfico da Figura 3.8 ilustra os resíduos standardizados em função da alavancagem bem como a distância de Cook, para cada um dos valores observados da variável resposta, destacando os valores y que têm maior efeito sobre as estimativas do parâmetro.

Pode observar-se que o ponto número 4111 tem uma alavancagem pouco superior ao ponto número 7577 e que ambas têm uma influência quase igual pois estão quase à mesma distância do contorno da distância de Cook. No entanto, para o ponto número 8718 observa-se que, em relação aos outros dois pontos influentes, têm uma alavancagem e uma influência muito inferior.

Uma forma de obter o gráfico da Figura 3.9 é recorrer à função `rstandard` do Software *R* que estima os resíduos *standard* quando se tem uma variável dependente apenas com duas categorias, como no caso da variável *Default*. Os valores dos resíduos, para cada categoria da variável resposta, serão iguais em magnitude e opostos em sinal para cada par de valores na mesma categoria, como reflexo das Equações (3.24) e (3.25).

Para investigar o quanto estes pontos influentes afectam as estimativas dos parâ-

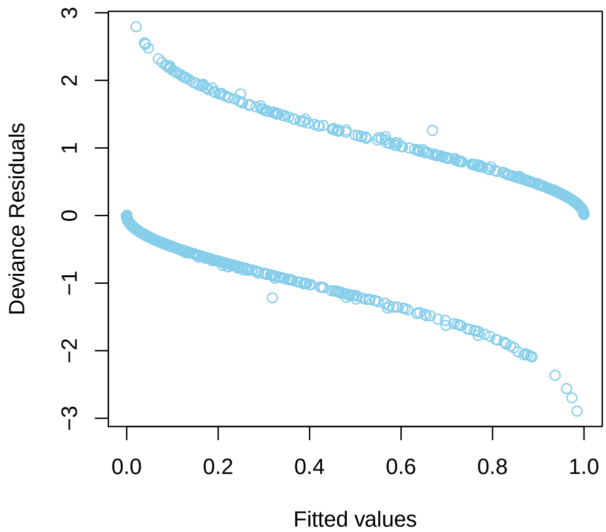

Figura 3.9: Desvios Residuais × Valores Ajustados

metros e os seus erros, dever-se-á repetir o processo de modelação estatística, e excluir os pontos em questão. Alternativamente poder-se-ia fazer um "*jackknife*" aos dados (veja-se [Crawley, 2007]), ou seja, excluir da carteira de crédito um cliente de cada vez e, em cada uma dessas exclusões, reestimar o parâmetro de interesse. Embora este tipo de análise não seja feita neste trabalho, pode dizer-se que o modelo estimado não é um bom modelo para os três pontos identificados na Figura (3.8).

Conclusão

Ao longo desta dissertação, o objectivo foi sempre o da estimação da probabilidade de *default*, com base nas características facultadas pelo solicitante de crédito.

Começou-se por uma abordagem geral de conceitos e métodos de avaliação de risco de crédito, fundamentais para o estudo de uma carteira de crédito.

No Capítulo 2 obteve-se um resultado importante para a estimação do *spread* mínimo a atribuir a cada cliente. Este resultado sugere que o *spread* mínimo a atribuir seja função da probabilidade de *default* e da taxa de recuperação.

Neste trabalho, foi desenvolvido um modelo para estimar a probabilidade de *default* e deixou-se para trabalhos posteriores a estimação da probabilidade de recuperação, no entanto, vale a pena ressaltar que, utilizando as variáveis adequadas, a estimação da probabilidade de recuperação faz-se de forma análoga à estimação da probabilidade de *default*.

Após o estudo dos modelos utilizados comummente na análise de risco de crédito, optou-se por utilizar o modelo de Credit Scoring de aprovação de crédito. Para analisar a carteira de crédito, utilizou-se ainda a técnica estatística de regressão linear múltipla, designadamente o modelo *Logit* para o cálculo da probabilidade e as técnicas *Stepwise-Backward* e *AIC* para a selecção das variáveis mais significativas no que respeita a explicar a ocorrência de *default*.

Ambas as técnicas de selecção aplicadas, Stepwise e AIC, revelaram que as variáveis que melhor explicam a ocorrência de *default*, de entre todas as que foram analisadas, são a Idade, o Nº de Dependentes, o Estado Civil, o Código Postal, o Rendimento, o Valor do Crédito, a Percentagem de Entrada, o Nº de Prestações Totais e o Nº de Prestações Pagas.

Relembrando os valores atribuídos aos pesos das variáveis (ver Tabela 3.3) no processo de simulação da variável Default concluiu-se que, as variáveis que os métodos de selecção consideram mais significativas, correspondem às variáveis cujos pesos atribuídos são mais elevados, variando entre os 5% e os 24%, evidenciando-se assim, a potencialidade destes métodos de selecção. A capacidade de selecção destes métodos é ainda verificada, quando a variável Pagas é seleccionada como significativa, mesmo não tendo sido associado um peso a esta variável em particular, mas sim à variável Pagas/Total, ou seja, o método é suficientemente robusto para identificar o peso que a variável Pagas tem no modelo pelo facto de estar associada a outra variável.

Da implementação do modelo Logit obtiveram-se os coeficientes, $\beta_i's$, que reflectem a capacidade preditora das respectivas variáveis em relação à variável Default. Deste modo, pode dizer-se que os coeficientes com valores negativos correspondem às variáveis que menos contribuem para a ocorrência de default, ao contrário dos coeficientes com valores positivos. Este facto advém da escolha feita inicialmente, para os valores da variável Default: valor 0 para clientes cumpridores e valor 1 para clientes incumpridores.

Para avaliar a capacidade do modelo prever a ocorrência de *default* estimaram-se as probabilidades de *default* de todos os clientes da carteira através do modelo Logit e, para cada cliente, foram comparados os resultados com os valores da variável Default simulada previamente.

Os resultados mostram que o modelo Logit consegue prever, em média, acertadamente cerca de $82,53\%$ dos casos, quando se define, para o modelo Logit, que as probabilidades acima de 50% são incumprimentos e prevê, em média, correctamente $88,63\%$ quando se considera que as probabilidades estimadas acima dos 25%, são consideradas situações em que ocorre *default*. Desta forma, tendo em consideração a elevada taxa de previsibilidade do modelo ajustado e os valores elevados do Coeficiente de Determinação Ajustado, considera-se que este modelo poderá ser utilizado para estimar a probabilidade de *default* para novos clientes solicitantes de um crédito à Habitação.

Ficam em aberto algumas questões, que não foram analisadas neste trabalho, mas que podem ser objecto de estudos futuros, nomeadamente:

- A Estimação da taxa de recuperação, necessária para o cálculo do spread mínimo a aplicar, mas que se deduz ser estimada de forma análoga à probabilidade de *default* recorrendo, no entanto, a variáveis explicativas distintas.

- Aplicação do modelo Vórtices Estocásticos, ver [Guerreiro e Mexia, 2008], a uma carteira de crédito dinâmica, isto é, agrupar os clientes em "classes de risco" (com base na probabilidade de *default* estimada), estimar as probabilidades de transição entre classes de risco e, assumindo que uma carteira de crédito aberta (permitindo a subscrição de novos contratos e a anulação de contratos existentes) analisar, numa perspectiva temporal , a evolução da dimensão de cada uma das classes de risco, estimando o peso relativo de cada uma das classes dentro da carteira de crédito.

Apêndice A

Tabelas

Neste apêndice apresentam-se algumas tabelas referidas ao longo deste trabalho, designadamente a Tabela da Moody's:"Projeção Global de Probabilidades de Transição a 1-Ano(Percentagem)", a Tabela das variáveis explicativas e respectivas categorias utilizadas na simulação da carteira de crédito e a Tabela que ilustra uma parte da carteira de crédito simulada.

	Aaa	Aa1	Aa2	Aa3	A1	A2	A3	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa1	Caa2	Caa3	Ca-C	DEF
Aaa	**92.4**	2.2	1.3	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aa1	4.3	**84.4**	4.6	2.0	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aa2	21.4	5.7	**78.4**	5.8	1.2	0.4	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aa3	30.1	1.0	4.1	**83.6**	3.5	1.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A1	10.0	0.1	0.4	5.7	**82.9**	4.1	1.7	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A2	20.0	0.0	0.2	1.0	5.0	**82.9**	4.1	1.4	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A3	30.1	0.1	0.1	0.2	0.9	7.6	**78.7**	4.5	2.1	0.7	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baa1	10.1	0.1	0.1	0.1	0.1	0.7	5.9	**79.8**	5.0	1.7	0.4	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Baa2	20.1	0.1	0.1	0.1	0.1	0.7	1.6	5.5	**78.6**	5.1	1.2	0.7	0.4	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0
Baa3	3.1	0.0	0.0	0.0	0.2	0.5	0.6	2.2	8.3	**76.2**	3.3	1.9	1.0	0.4	0.2	0.1	0.1	0.0	0.0	0.0	0.2
Ba1	10.0	0.0	0.1	0.1	0.5	0.4	0.4	1.2	4.1	13.8	**59.8**	3.9	3.4	1.3	0.9	0.5	0.2	0.1	0.0	0.1	0.2
Ba2	20.0	0.0	0.1	0.1	0.2	0.3	0.2	0.5	1.2	4.5	11.2	**58.6**	4.8	2.9	2.1	1.2	0.7	0.2	0.1	0.2	0.5
Ba3	30.0	0.0	0.1	0.1	0.0	0.3	0.2	0.2	0.6	1.2	3.8	9.5	**60.6**	4.1	4.3	1.9	0.7	0.4	0.2	0.1	0.5
B1	10.0	0.0	0.0	0.0	0.1	0.2	0.1	0.3	0.2	0.4	1.0	2.9	8.5	**63.1**	5.4	4.1	1.6	0.9	0.3	0.2	0.8
B2	20.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.2	0.2	0.4	0.8	2.5	7.8	**65.8**	5.3	3.0	1.6	0.6	0.4	1.6
B3	30.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.3	0.5	3.0	6.2	**64.9**	4.1	3.6	1.0	1.1	3.1
Caa1	10.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.2	0.2	0.7	2.3	5.8	**55.0**	4.5	3.6	2.5	7.7
Caa2	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.2	0.0	0.5	0.6	0.8	2.8	3.4	**51.4**	2.3	4.2	14.0
Caa3	30.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.0	0.1	1.4	0.5	1.4	3.3	4.8	**40.1**	5.7	18.3
Ca-C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.3	0.7	0.4	1.2	2.2	**35.3**	24.4

Tabela A.1: Moody’s:Projeção Global de Probabilidades de Transição a 1-Ano(%)

Variáveis	Categorias	Descrição	%
Idade (anos)	Jovens	[18,25]	13,00%
	Adultos	[26,50]	85,00%
	Idosos	>50	2,00%
Nº Dependentes	Poucos	0 ou 1	80,00%
	Alguns	2 ou 3	15,00%
	Muitos	$\geq$ 4	5,00%
Tipo Habitação	Casa	Moradia, Apartamento	89,00%
	Empresa	Prédio, Empresa	3,00%
	Comercio	Comércio, Terreno, Outro	8,00%
Idade da Casa	Nova	$\leq$ 5 anos	75,00%
	Usada	[6,15] anos	20,00%
	Velha	> 15	5,00%
Estado Civil	Div	Divorciado, Separado	2,60%
	Sol	Solteiro, União Livre, Viúvo	47,80%
	Cas	Casado	49,60%
Códigos Postais	Zona 1	11, 17, 20, 21, 25, 33, 38, 41, 42, 43, 53, 60, 62, 72, 77, 79, 8, 83, 85, 92, 97, 98	65,0%
	Zona 2	13, 22, 24, 26, 27, 29, 31, 49, 73, 96	20,00%
	Zona 3	12, 18, 19, 23, 28, 34, 44, 46, 47, 70, 74, 90	10,00%
	Zona 4	10, 14, 15, 16, 30, 32, 35, 36, 37, 40, 45, 48, 50, 51, 52, 54, 61, 63, 64, 71, 75, 76, 78, 80, 82, 86, 87, 88, 89, 91, 93, 94, 95, 99	5,00%
Tipo Profissão	Tipo 1	Agricultor, ANI/Empresário, Comerciante, Gerente de Sociedade, Profissão Liberal, Artista, Jornalista, Polícia, Militar, GNR, Camionista, Taxista, Pescador, Assalariado Agrícola, Desempregado, Vive de Rendimentos, Estudante, Doméstica, Trabalhador Temporário	20,0%
	Tipo 2	Administrativo, Auxiliar Função Pública, Empregado Escritório, Empregado Comércio, Empregado Serviços, Operário, Outros	30,00%
	Tipo 3	Professor Primário, Enfermeira, Assistente Social, Parteira, Quadro Médio Função Pública, Quadro Médio Sector Privado, Técnicos, Chefe de equipa, Contramestre	35,00%
	Tipo 4	Quadro Superior da Função Pública, Professor, Investigador, Médico, Quandro Superior Sector Privado, Reformado	15,00%
Antiguidade Profissão (anos)	Recente	$\leq$ 1	7,50%
	Antigo	>1	92,50%
Rendimento Líquido (€)	Minimo	$\leq$ 750	40,0%
	Baixo	]750,1500]	37,00%
	Médio	]1500,5000]	20,00%
	Alto	> 5000	3,00%
Crédito(€)			
Percentagem de Entrada	Menos10	$\leq$ 10	-
	Mais10	>10	-
Nº Total Prestações (meses)	1	60	0,16%
	2	120	1,06%
	3	180	4,25%
	4	240	11,15%
	5	300	20,07%
	6	360	25,08%
	7	420	21,50%
	8	480	12,09%
	9	540	4,03%
	10	600	0,60%
Pagas		Nº Aleatório entre 60 e Nº Total Prestações	
Pagas / Total (%)	1	$\leq$ 25	-
	2	> 25 e $\leq$ 50	-
	3	> 50 e $\leq$ 75	-
	4	> 75 e $\leq$ 100	-

Tabela A.2: Variáveis Explicativas e Respectivas Categorias

Cliente	Idade	Dep	Hab	Idcasa	Civil	Postal	Prof	Antg	Rend	Cred	Entr	Ptotal	Pagas	Decpercnt	Default
1	Adultos	Poucos	Comercio	Nova	Cas	Zona4	Tipo2	Recente	Medio	159216	Mais10	420	94	MenosEntre25e505	0
2	Adultos	Poucos	Casa	Nova	Sol	Zona4	Tipo3	Antigo	Minimo	96878	Menos10	360	160	Entre25e50	1
3	Adultos	Poucos	Casa	Nova	Cas	Zona4	Tipo3	Antigo	Minimo	121671	Menos10	240	164	Entre50e75	0
4	Adultos	Poucos	Casa	Velha	Sol	Zona3	Tipo4	Antigo	Minimo	211386	Mais10	420	300	Entre50e75	0
5	Adultos	Poucos	Casa	Velha	Sol	Zona2	Tipo4	Antigo	Baixo	281702	Menos10	300	97	Entre25e50	0
6	Adultos	Poucos	Casa	Nova	Sol	Zona3	Tipo2	Antigo	Baixo	229578	Menos10	240	175	Entre50e75	0
7	Adultos	Poucos	Casa	Nova	Cas	Zona4	Tipo2	Recente	Minimo	92500	Menos10	360	200	Entre50e75	0
8	Adultos	Poucos	Casa	Nova	Cas	Zona3	Tipo2	Antigo	Minimo	228955	Mais10	420	287	Entre50e75	0
9	Adultos	Poucos	Casa	Nova	Sol	Zona2	Tipo2	Antigo	Medio	141543	Menos10	420	192	Entre25e50	0
10	Adultos	Poucos	Casa	Usada	Cas	Zona4	Tipo2	Recente	Minimo	194228	Mais10	300	88	Entre25e50	0
11	Adultos	Poucos	Casa	Nova	Cas	Zona4	Tipo3	Antigo	Baixo	158808	Menos10	480	155	Entre25e50	0
12	Adultos	Poucos	Casa	Nova	Sol	Zona4	Tipo2	Antigo	Minimo	166335	Menos10	300	110	Entre25e50	0
13	Adultos	Poucos	Casa	Nova	Sol	Zona1	Tipo3	Antigo	Alto	98277	Mais10	360	331	Mais75	0
14	Adultos	Poucos	Casa	Nova	Sol	Zona4	Tipo3	Antigo	Minimo	198219	Menos10	420	276	Entre50e75	0
15	Adultos	Poucos	Casa	Nova	Sol	Zona4	Tipo4	Antigo	Medio	121409	Menos10	300	177	Entre50e75	0
16	Adultos	Poucos	Casa	Nova	Cas	Zona3	Tipo1	Antigo	Minimo	397506	Menos10	360	238	Entre50e75	0
17	Adultos	Poucos	Casa	Nova	Cas	Zona4	Tipo1	Antigo	Minimo	101402	Mais10	360	171	Entre25e50	0
18	Adultos	Poucos	Casa	Nova	Cas	Zona4	Tipo1	Antigo	Baixo	121999	Mais10	300	287	Mais75	0
19	Adultos	Poucos	Casa	Nova	Cas	Zona4	Tipo3	Antigo	Minimo	111252	Menos10	420	286	Entre50e75	0
20	Adultos	Poucos	Casa	Nova	Cas	Zona3	Tipo4	Antigo	Medio	167151	Mais10	480	403	Mais75	0
21	Adultos	Poucos	Comercio	Usada	Cas	Zona4	Tipo3	Antigo	Minimo	160103	Mais10	240	148	Entre50e75	0
22	Adultos	Poucos	Casa	Nova	Cas	Zona1	Tipo3	Antigo	Baixo	389209	Mais10	420	305	Entre50e75	0
23	Adultos	Alguns	Casa	Nova	Cas	Zona4	Tipo2	Antigo	Minimo	322003	Menos10	360	86	MenosEntre25e505	0
24	Adultos	Poucos	Empresa	Nova	Cas	Zona4	Tipo3	Antigo	Baixo	134882	Menos10	480	324	Entre50e75	0
25	Adultos	Poucos	Comercio	Nova	Sol	Zona4	Tipo4	Antigo	Medio	249243	Menos10	480	182	Entre25e50	0
26	Adultos	Poucos	Casa	Nova	Cas	Zona4	Tipo3	Antigo	Minimo	175836	Mais10	480	241	Entre50e75	0
27	Adultos	Poucos	Casa	Nova	Sol	Zona4	Tipo1	Antigo	Medio	90976	Mais10	360	196	Entre50e75	0
28	Adultos	Poucos	Casa	Nova	Cas	Zona4	Tipo3	Antigo	Medio	96701	Mais10	360	86	MenosEntre25e505	0
29	Adultos	Alguns	Casa	Nova	Sol	Zona4	Tipo1	Antigo	Baixo	109513	Mais10	360	61	MenosEntre25e505	0
30	Adultos	Poucos	Casa	Nova	Sol	Zona4	Tipo3	Antigo	Baixo	199968	Mais10	420	206	Entre25e50	0

Tabela A.3: Carteira de Crédito - Base de Dados

Apêndice B

Inputs e Outputs do R

Neste apêndice apresentam-se alguns inputs utilizados e os outputs gerados(deve ter-se em consideração que os inputs apresentados são apenas exemplificativos, e o autor não se responsabiliza por qualquer utilização indevida dos mesmos).

Para 4 Zonas Postais

```
> data <- read.table("DADOSCRz1z2semDec.txt", header = TRUE)
> attach(data)
> names(data)
```

```
 [1]  "Idade"     "Dep"       "Hab"       "Idcasa"    "Civil"
 [6]  "Postal"    "Prof"      "Antg"      "Rend"      "Cred"
 [11] "Entr"      "Ptotal"    "Pagas"     "Pagpercnt" "Default"
```

»»

Regressão Logística

➔Modelo Completo

```
> Modfull
```

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.92e+00  -7.13e-03  -3.79e-04  -1.29e-05   2.80e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.59e+01   6.07e+02   -0.08    0.940
IdadeIdosos -2.09e+00   8.74e-01   -2.40    0.017 *
IdadeJovens  3.95e+00   3.22e-01   12.29  < 2e-16 ***
DepMuitos    1.30e+00   6.89e-01    1.89    0.059 .
DepPoucos    2.85e+00   3.53e-01    8.05  8.1e-16 ***
HabComercio  1.26e-02   3.68e-01    0.03    0.973
HabEmpresa  -2.70e-01   6.09e-01   -0.44    0.657
IdcasaUsada -6.73e-02   2.61e-01   -0.26    0.796
IdcasaVelha  5.45e-01   4.82e-01    1.13    0.259
CivilDiv     3.38e+00   6.04e-01    5.60  2.1e-08 ***
CivilSol     1.60e+00   2.28e-01    6.99  2.7e-12 ***
PostalZona2  2.35e+00   1.49e+00    1.58    0.115
```

```
PostalZona3  5.48e+00   1.41e+00    3.90  9.8e-05 ***
PostalZona4  7.69e+00   1.42e+00    5.41  6.3e-08 ***
ProfTipo2   -3.32e-01   2.98e-01   -1.12    0.265
ProfTipo3   -5.11e-01   2.85e-01   -1.79    0.073 .
ProfTipo4   -6.36e-01   3.63e-01   -1.75    0.080 .
AntgRecente -5.28e-01   3.84e-01   -1.38    0.169
RendBaixo    1.72e+01   6.07e+02    0.03    0.977
RendMedio    8.74e+00   6.07e+02    0.01    0.989
RendMinimo   2.38e+01   6.07e+02    0.04    0.969
Cred        -1.06e-05   1.67e-06   -6.37  1.9e-10 ***
EntrMenos10  5.14e+00   3.75e-01   13.70  < 2e-16 ***
Ptotal       3.66e-02   2.28e-03   16.06  < 2e-16 ***
Pagas       -5.70e-02   3.65e-03  -15.62  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  644.28  on 9975  degrees of freedom
AIC: 694.3

Number of Fisher Scoring iterations: 19
```

➔Modelo Nulo

```
> Modnull


Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.258  -0.258  -0.258  -0.258   2.615

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3871     0.0562   -60.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance: 2880.1  on 9999  degrees of freedom
AIC: 2882

Number of Fisher Scoring iterations: 6
```

➔Comparação Modelo Nulo Vs Modelo Completo

```
> anova(Modnull, Modfull, test = "Chisq")
Analysis of Deviance Table

Model 1: Default ~ 1
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9999       2880
2      9975        644 24     2236    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

»»

Os resultados sugerem a junção da Zona 1 com a Zona 2 da variável Código Postal.

Para 3 Zonas Postais

```
> data <- read.table("DADOSCRz1z3semDec.txt", header = TRUE)
> attach(data)
> names(data)
```

```
 [1]  "Idade"     "Dep"       "Hab"       "Idcasa"    "Civil"
 [6]  "Postal"    "Prof"      "Antg"      "Rend"      "Cred"
 [11] "Entr"      "Ptotal"    "Pagas"     "Pagpercnt" "Default"
```

Regressão Logística

➔Modelo Completo

```
> Modfull
```

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.91e+00  -7.29e-03  -4.02e-04  -1.40e-05   2.79e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.39e+01   6.07e+02   -0.07    0.942
IdadeIdosos -2.07e+00   8.73e-01   -2.37    0.018 *
IdadeJovens  3.92e+00   3.18e-01   12.31  < 2e-16 ***
DepMuitos    1.30e+00   6.88e-01    1.89    0.058 .
DepPoucos    2.84e+00   3.53e-01    8.06  7.9e-16 ***
HabComercio  1.26e-02   3.68e-01    0.03    0.973
HabEmpresa  -3.23e-01   5.99e-01   -0.54    0.590
IdcasaUsada -6.92e-02   2.60e-01   -0.27    0.790
IdcasaVelha  4.79e-01   4.90e-01    0.98    0.328
CivilDiv     3.38e+00   6.00e-01    5.63  1.8e-08 ***
CivilSol     1.60e+00   2.28e-01    7.02  2.3e-12 ***
PostalZona3  3.58e+00   6.57e-01    5.45  5.1e-08 ***
PostalZona4  5.78e+00   6.65e-01    8.69  < 2e-16 ***
ProfTipo2   -3.72e-01   2.96e-01   -1.26    0.209
ProfTipo3   -5.49e-01   2.84e-01   -1.94    0.053 .
ProfTipo4   -6.62e-01   3.62e-01   -1.83    0.068 .
AntgRecente -5.21e-01   3.81e-01   -1.37    0.172
RendBaixo    1.72e+01   6.07e+02    0.03    0.977
RendMedio    8.79e+00   6.07e+02    0.01    0.988
RendMinimo   2.37e+01   6.07e+02    0.04    0.969
Cred        -1.05e-05   1.66e-06   -6.34  2.3e-10 ***
EntrMenos10  5.11e+00   3.73e-01   13.71  < 2e-16 ***
Ptotal       3.65e-02   2.27e-03   16.11  < 2e-16 ***
Pagas       -5.67e-02   3.62e-03  -15.67  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance:  646.9  on 9976  degrees of freedom
AIC: 694.9

Number of Fisher Scoring iterations: 19
```

➔Modelo Nulo

> Modnull

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.258  -0.258  -0.258  -0.258   2.615

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3871     0.0562   -60.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance: 2880.1  on 9999  degrees of freedom
AIC: 2882

Number of Fisher Scoring iterations: 6
```

➔ Comparação Modelo Nulo Vs Modelo Completo

> anova(Modnull, Modfull, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ 1
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9999       2880
2      9976        647 23     2233    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> R2_full

```
[1] 0.775
```

»»

Stepwise-Backward

> a_1 <- (Modfull - Rend)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1482  -0.1194  -0.0395  -0.0123   2.8155

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17e+01   7.03e-01  -16.70  < 2e-16 ***
IdadeIdosos -7.57e-01   6.58e-01   -1.15     0.25
IdadeJovens  1.79e+00   1.75e-01   10.20  < 2e-16 ***
DepMuitos    6.92e-01   4.96e-01    1.39     0.16
DepPoucos    1.29e+00   2.38e-01    5.41  6.2e-08 ***
HabComercio  1.17e-01   2.38e-01    0.49     0.62
HabEmpresa   1.36e-02   3.81e-01    0.04     0.97
IdcasaUsada  1.15e-01   1.67e-01    0.69     0.49
IdcasaVelha  1.26e-01   3.35e-01    0.38     0.71
```

```
CivilDiv     1.51e+00   3.73e-01    4.06  4.9e-05 ***
CivilSol     7.72e-01   1.46e-01    5.28  1.3e-07 ***
PostalZona3  1.92e+00   3.98e-01    4.83  1.4e-06 ***
PostalZona4  2.75e+00   3.74e-01    7.35  2.0e-13 ***
ProfTipo2   -2.27e-01   1.97e-01   -1.15     0.25
ProfTipo3   -1.84e-01   1.92e-01   -0.96     0.34
ProfTipo4   -2.08e-01   2.37e-01   -0.88     0.38
AntgRecente -3.60e-01   2.81e-01   -1.28     0.20
Cred        -4.94e-06   1.05e-06   -4.71  2.5e-06 ***
EntrMenos10  2.66e+00   2.13e-01   12.51  < 2e-16 ***
Ptotal       1.78e-02   9.74e-04   18.32  < 2e-16 ***
Pagas       -2.85e-02   1.62e-03  -17.58  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance: 1483.5  on 9979  degrees of freedom
AIC: 1526

Number of Fisher Scoring iterations: 9
```

> anova(a_1, Modfull, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Cred + Entr + Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9979       1484
2      9976        647  3      837    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> a_2 <- (Modfull - Hab)

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.908384  -0.007320  -0.000400  -0.000014   2.796239

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.39e+01   6.07e+02   -0.07    0.942
IdadeIdosos -2.09e+00   8.70e-01   -2.40    0.016 *
IdadeJovens  3.93e+00   3.18e-01   12.35  < 2e-16 ***
DepMuitos    1.32e+00   6.86e-01    1.92    0.055 .
DepPoucos    2.84e+00   3.53e-01    8.06  7.5e-16 ***
IdcasaUsada -7.74e-02   2.59e-01   -0.30    0.765
IdcasaVelha  4.56e-01   4.86e-01    0.94    0.348
CivilDiv     3.38e+00   6.00e-01    5.64  1.7e-08 ***
CivilSol     1.60e+00   2.28e-01    7.01  2.4e-12 ***
PostalZona3  3.59e+00   6.55e-01    5.49  4.0e-08 ***
PostalZona4  5.79e+00   6.62e-01    8.75  < 2e-16 ***
ProfTipo2   -3.71e-01   2.96e-01   -1.25    0.210
ProfTipo3   -5.54e-01   2.83e-01   -1.96    0.050 .
ProfTipo4   -6.65e-01   3.62e-01   -1.84    0.066 .
AntgRecente -5.19e-01   3.81e-01   -1.36    0.173
RendBaixo    1.72e+01   6.07e+02    0.03    0.977
RendMedio    8.78e+00   6.07e+02    0.01    0.988
RendMinimo   2.37e+01   6.07e+02    0.04    0.969
```

```
Cred        -1.05e-05   1.65e-06   -6.37  1.9e-10 ***
EntrMenos10  5.11e+00   3.73e-01   13.69  < 2e-16 ***
Ptotal       3.65e-02   2.26e-03   16.11  < 2e-16 ***
Pagas       -5.67e-02   3.61e-03  -15.69  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance:  647.2  on 9978  degrees of freedom
AIC: 691.2

Number of Fisher Scoring iterations: 19
```

> anova(a_2, Modfull, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Idcasa + Civil + Postal + Prof + Antg +
    Rend + Cred + Entr + Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9978        647
2      9976        647  2    0.302      0.86
```

> a_3 <- (a_2- Idcasa)

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.93e+00  -7.49e-03  -4.15e-04  -1.45e-05   2.79e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.38e+01   6.09e+02   -0.07    0.943
IdadeIdosos -2.05e+00   8.72e-01   -2.35    0.019 *
IdadeJovens  3.92e+00   3.17e-01   12.35  < 2e-16 ***
DepMuitos    1.36e+00   6.87e-01    1.97    0.048 *
DepPoucos    2.85e+00   3.52e-01    8.08  6.3e-16 ***
CivilDiv     3.36e+00   6.00e-01    5.61  2.1e-08 ***
CivilSol     1.58e+00   2.27e-01    6.97  3.1e-12 ***
PostalZona3  3.56e+00   6.51e-01    5.47  4.4e-08 ***
PostalZona4  5.74e+00   6.58e-01    8.74  < 2e-16 ***
ProfTipo2   -3.64e-01   2.95e-01   -1.23    0.217
ProfTipo3   -5.62e-01   2.83e-01   -1.99    0.047 *
ProfTipo4   -6.76e-01   3.62e-01   -1.87    0.062 .
AntgRecente -5.11e-01   3.81e-01   -1.34    0.180
RendBaixo    1.72e+01   6.09e+02    0.03    0.977
RendMedio    8.78e+00   6.09e+02    0.01    0.989
RendMinimo   2.37e+01   6.09e+02    0.04    0.969
Cred        -1.05e-05   1.65e-06   -6.36  2.1e-10 ***
EntrMenos10  5.10e+00   3.72e-01   13.71  < 2e-16 ***
Ptotal       3.64e-02   2.25e-03   16.14  < 2e-16 ***
Pagas       -5.64e-02   3.59e-03  -15.71  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance:  648.2  on 9980  degrees of freedom
AIC: 688.2

Number of Fisher Scoring iterations: 19
```

> anova(a_3, a_2, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Prof + Antg + Rend +
    Cred + Entr + Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Idcasa + Civil + Postal + Prof + Antg +
    Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9980        648
2      9978        647  2     1.01       0.6
```

> a_4 <- (a_3- Prof)

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.91e+00  -7.91e-03  -4.47e-04  -1.63e-05   2.79e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.38e+01   6.14e+02   -0.07    0.943
IdadeIdosos -2.19e+00   8.72e-01   -2.51    0.012 *
IdadeJovens  3.86e+00   3.12e-01   12.37  < 2e-16 ***
DepMuitos    1.26e+00   6.91e-01    1.82    0.068 .
DepPoucos    2.79e+00   3.49e-01    7.99  1.3e-15 ***
CivilDiv     3.26e+00   5.93e-01    5.50  3.8e-08 ***
CivilSol     1.55e+00   2.25e-01    6.88  5.9e-12 ***
PostalZona3  3.42e+00   6.31e-01    5.42  6.0e-08 ***
PostalZona4  5.59e+00   6.36e-01    8.79  < 2e-16 ***
AntgRecente -4.84e-01   3.82e-01   -1.27    0.205
RendBaixo    1.71e+01   6.14e+02    0.03    0.978
RendMedio    8.71e+00   6.14e+02    0.01    0.989
RendMinimo   2.35e+01   6.14e+02    0.04    0.969
Cred        -1.01e-05   1.63e-06   -6.20  5.6e-10 ***
EntrMenos10  5.10e+00   3.72e-01   13.71  < 2e-16 ***
Ptotal       3.59e-02   2.21e-03   16.24  < 2e-16 ***
Pagas       -5.58e-02   3.53e-03  -15.81  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  653.19  on 9983  degrees of freedom
AIC: 687.2

Number of Fisher Scoring iterations: 19
```

> anova(a_4, a_3, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Antg + Rend + Cred +
    Entr + Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Civil + Postal + Prof + Antg + Rend +
    Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9983        653
2      9980        648  3     4.98      0.17
```

> a_5 <- (a_4- Antg)

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.89e+00  -7.92e-03  -4.51e-04  -1.67e-05   2.79e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37e+01   6.14e+02   -0.07    0.943
IdadeIdosos -2.15e+00   8.71e-01   -2.46    0.014 *
IdadeJovens  3.85e+00   3.12e-01   12.35  < 2e-16 ***
DepMuitos    1.27e+00   6.86e-01    1.85    0.065 .
DepPoucos    2.77e+00   3.48e-01    7.97  1.6e-15 ***
CivilDiv     3.22e+00   6.00e-01    5.36  8.1e-08 ***
CivilSol     1.56e+00   2.25e-01    6.93  4.3e-12 ***
PostalZona3  3.45e+00   6.30e-01    5.48  4.3e-08 ***
PostalZona4  5.61e+00   6.36e-01    8.82  < 2e-16 ***
RendBaixo    1.71e+01   6.14e+02    0.03    0.978
RendMedio    8.76e+00   6.14e+02    0.01    0.989
RendMinimo   2.35e+01   6.14e+02    0.04    0.969
Cred        -1.02e-05   1.63e-06   -6.28  3.3e-10 ***
EntrMenos10  5.07e+00   3.69e-01   13.73  < 2e-16 ***
Ptotal       3.60e-02   2.22e-03   16.23  < 2e-16 ***
Pagas       -5.57e-02   3.52e-03  -15.81  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  654.86  on 9984  degrees of freedom
AIC: 686.9

Number of Fisher Scoring iterations: 19
```

> anova(a_5, a_4, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Civil + Postal + Antg + Rend + Cred +
    Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9984        655
2      9983        653  1     1.67      0.20
```

➔Melhor Modelo Estimado com Stepwise

> Mmod_Stp

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.89e+00  -7.92e-03  -4.51e-04  -1.67e-05   2.79e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37e+01   6.14e+02   -0.07    0.943
IdadeIdosos -2.15e+00   8.71e-01   -2.46    0.014 *
IdadeJovens  3.85e+00   3.12e-01   12.35  < 2e-16 ***
DepMuitos    1.27e+00   6.86e-01    1.85    0.065 .
DepPoucos    2.77e+00   3.48e-01    7.97  1.6e-15 ***
CivilDiv     3.22e+00   6.00e-01    5.36  8.1e-08 ***
```

```
CivilSol     1.56e+00   2.25e-01    6.93  4.3e-12 ***
PostalZona3  3.45e+00   6.30e-01    5.48  4.3e-08 ***
PostalZona4  5.61e+00   6.36e-01    8.82  < 2e-16 ***
Cred        -1.02e-05   1.63e-06   -6.28  3.3e-10 ***
EntrMenos10  5.07e+00   3.69e-01   13.73  < 2e-16 ***
Ptotal       3.60e-02   2.22e-03   16.23  < 2e-16 ***
Pagas       -5.57e-02   3.52e-03  -15.81  < 2e-16 ***
RendBaixo    1.71e+01   6.14e+02    0.03    0.978
RendMedio    8.76e+00   6.14e+02    0.01    0.989
RendMinimo   2.35e+01   6.14e+02    0.04    0.969
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  654.86  on 9984  degrees of freedom
AIC: 686.9

Number of Fisher Scoring iterations: 19
```

»»

Step AIC

```
Start:  AIC=695
Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas

         Df Deviance  AIC
- Hab     2      647  691
- Idcasa  2      648  692
- Prof    3      652  694
- Antg    1      649  695
<none>           647  695
- Cred    1      694  740
- Civil   2      719  763
- Dep     2      737  781
- Postal  2      847  891
- Idade   2      861  905
- Entr    1     1049 1095
- Ptotal  1     1417 1463
- Rend    3     1484 1526
- Pagas   1     1791 1837

Step:  AIC=691
Default ~ Idade + Dep + Idcasa + Civil + Postal + Prof + Antg +
    Rend + Cred + Entr + Ptotal + Pagas

         Df Deviance  AIC
- Idcasa  2      648  688
- Prof    3      652  690
- Antg    1      649  691
<none>           647  691
+ Hab     2      647  695
- Cred    1      695  737
- Civil   2      719  759
- Dep     2      737  777
- Postal  2      848  888
- Idade   2      863  903
- Entr    1     1049 1091
- Ptotal  1     1419 1461
- Rend    3     1484 1522
- Pagas   1     1793 1835
```

```
Step:  AIC=688
Default ~ Idade + Dep + Civil + Postal + Prof + Antg + Rend +
    Cred + Entr + Ptotal + Pagas

         Df Deviance  AIC
- Prof    3      653  687
- Antg    1      650  688
<none>           648  688
+ Idcasa  2      647  691
+ Hab     2      648  692
- Cred    1      696  734
- Civil   2      720  756
- Dep     2      739  775
- Postal  2      848  884
- Idade   2      864  900
- Entr    1     1050 1088
- Ptotal  1     1419 1457
- Rend    3     1484 1518
- Pagas   1     1796 1834

Step:  AIC=687
Default ~ Idade + Dep + Civil + Postal + Antg + Rend + Cred +
    Entr + Ptotal + Pagas

         Df Deviance  AIC
- Antg    1      655  687
<none>           653  687
+ Prof    3      648  688
+ Idcasa  2      652  690
+ Hab     2      653  691
- Cred    1      698  730
- Civil   2      722  752
- Dep     2      742  772
- Postal  2      850  880
- Idade   2      867  897
- Entr    1     1056 1088
- Ptotal  1     1420 1452
- Rend    3     1486 1514
- Pagas   1     1798 1830

Step:  AIC=687
Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas

         Df Deviance  AIC
<none>           655  687
+ Antg    1      653  687
+ Prof    3      650  688
+ Idcasa  2      654  690
+ Hab     2      655  691
- Cred    1      701  731
- Civil   2      724  752
- Dep     2      742  770
- Postal  2      852  880
- Idade   2      868  896
- Entr    1     1057 1087
- Ptotal  1     1422 1452
- Rend    3     1487 1513
- Pagas   1     1799 1829


Coefficients:
(Intercept)  IdadeIdosos  IdadeJovens    DepMuitos    DepPoucos     CivilDiv
  -4.37e+01    -2.15e+00     3.85e+00     1.27e+00     2.77e+00     3.22e+00
 CivilSol  PostalZona3  PostalZona4    RendBaixo    RendMedio
```

```
 1.56e+00     3.45e+00     5.61e+00     1.71e+01     8.76e+00
 RendMinimo         Cred  EntrMenos10       Ptotal        Pagas
   2.35e+01    -1.02e-05     5.07e+00     3.60e-02    -5.57e-02

Degrees of Freedom: 9999 Total (i.e. Null);  9984 Residual
Null Deviance:           2880
Residual Deviance: 655        AIC: 687
```

➔Melhor Modelo Estimado com Step AIC

> ModAIC

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.89e+00  -7.92e-03  -4.51e-04  -1.67e-05   2.79e+00

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37e+01   6.14e+02   -0.07    0.943
IdadeIdosos -2.15e+00   8.71e-01   -2.46    0.014 *
IdadeJovens  3.85e+00   3.12e-01   12.35  < 2e-16 ***
DepMuitos    1.27e+00   6.86e-01    1.85    0.065 .
DepPoucos    2.77e+00   3.48e-01    7.97  1.6e-15 ***
CivilDiv     3.22e+00   6.00e-01    5.36  8.1e-08 ***
CivilSol     1.56e+00   2.25e-01    6.93  4.3e-12 ***
PostalZona3  3.45e+00   6.30e-01    5.48  4.3e-08 ***
PostalZona4  5.61e+00   6.36e-01    8.82  < 2e-16 ***
RendBaixo    1.71e+01   6.14e+02    0.03    0.978
RendMedio    8.76e+00   6.14e+02    0.01    0.989
RendMinimo   2.35e+01   6.14e+02    0.04    0.969
Cred        -1.02e-05   1.63e-06   -6.28  3.3e-10 ***
EntrMenos10  5.07e+00   3.69e-01   13.73  < 2e-16 ***
Ptotal       3.60e-02   2.22e-03   16.23  < 2e-16 ***
Pagas       -5.57e-02   3.52e-03  -15.81  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  654.86  on 9984  degrees of freedom
AIC: 686.9

Number of Fisher Scoring iterations: 19
```

> R2_AIC

```
[1] 0.772
```

> anova(ModAIC, Modfull, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9984        655
2      9976        647  8     7.96      0.44
```

> ModAIC2 <- (ModAIC - Rend)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1117  -0.1210  -0.0397  -0.0124   2.7824

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.18e+01   6.90e-01  -17.16  < 2e-16 ***
IdadeIdosos -7.60e-01   6.58e-01   -1.15     0.25
IdadeJovens  1.78e+00   1.75e-01   10.16  < 2e-16 ***
DepMuitos    6.82e-01   4.93e-01    1.39     0.17
DepPoucos    1.27e+00   2.37e-01    5.35  8.8e-08 ***
CivilDiv     1.48e+00   3.72e-01    3.99  6.6e-05 ***
CivilSol     7.70e-01   1.46e-01    5.28  1.3e-07 ***
PostalZona3  1.91e+00   3.98e-01    4.79  1.7e-06 ***
PostalZona4  2.73e+00   3.73e-01    7.31  2.8e-13 ***
Cred        -4.92e-06   1.05e-06   -4.69  2.7e-06 ***
EntrMenos10  2.67e+00   2.12e-01   12.57  < 2e-16 ***
Ptotal       1.78e-02   9.72e-04   18.33  < 2e-16 ***
Pagas       -2.85e-02   1.62e-03  -17.60  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.1  on 9999  degrees of freedom
Residual deviance: 1487.4  on 9987  degrees of freedom
AIC: 1513

Number of Fisher Scoring iterations: 9
```

> anova(ModAIC2, ModAIC, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Cred + Entr + Ptotal +
    Pagas
Model 2: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9987       1487
2      9984        655  3      833    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> R2_AIC2

```
[1] 0.483
```

»»

Confimação Selecção variáveis

➔1ª função

> Modfull

```
Single term deletions

Model:
Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
```

```
    Antg + Rend + Cred + Entr + Ptotal + Pagas
       Df Deviance  AIC  LRT Pr(Chi)
<none>         647  695
Idade   2      861  905  215 < 2e-16 ***
Dep     2      737  781   90 < 2e-16 ***
Hab     2      647  691    0    0.86
Idcasa  2      648  692    1    0.59
Civil   2      719  763   72 < 2e-16 ***
Postal  2      847  891  200 < 2e-16 ***
Prof    3      652  694    5    0.19
Antg    1      649  695    2    0.16
Rend    3     1484 1526  837 < 2e-16 ***
Cred    1      694  740   48 5.5e-12 ***
Entr    1     1049 1095  402 < 2e-16 ***
Ptotal  1     1417 1463  770 < 2e-16 ***
Pagas   1     1791 1837 1144 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

>ModAIC

```
Single term deletions

Model:
Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
       Df Deviance  AIC  LRT Pr(Chi)
<none>         655  687
Idade   2      868  896  213 < 2e-16 ***
Dep     2      742  770   88 < 2e-16 ***
Civil   2      724  752   69 1.1e-15 ***
Postal  2      852  880  197 < 2e-16 ***
Rend    3     1487 1513  833 < 2e-16 ***
Cred    1      701  731   46 9.8e-12 ***
Entr    1     1057 1087  402 < 2e-16 ***
Ptotal  1     1422 1452  767 < 2e-16 ***
Pagas   1     1799 1829 1144 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➔2ª função

> Modfull

```
Single term additions

Model:
Default ~ Idade + Dep + Hab + Idcasa + Civil + Postal + Prof +
    Antg + Rend + Cred + Entr + Ptotal + Pagas
              Df Deviance AIC   LRT Pr(Chi)
<none>                647 695
Idade:Dep      3      646 700  1.22   0.749
Idade:Hab      4      642 698  4.68   0.322
Idade:Idcasa   4      645 701  1.82   0.769
Idade:Civil    4      646 702  1.13   0.889
Idade:Postal   4      644 700  3.07   0.546
Idade:Prof     6      645 705  2.15   0.906
Idade:Antg     2      646 698  0.40   0.817
Idade:Rend     6      647 707  0.29   1.000
Idade:Cred     2      640 692  6.88   0.032 *
Idade:Entr     2      646 698  0.58   0.746
Idade:Ptotal   2      646 698  0.86   0.650
Idade:Pagas    2      645 697  1.78   0.411
```

```
Dep:Hab        4      637 693  9.78     0.044 *
Dep:Idcasa     4      644 700  2.44     0.656
Dep:Civil      4      644 700  3.13     0.536
Dep:Postal     4      645 701  2.39     0.665
Dep:Prof       6      643 703  4.13     0.658
Dep:Antg       2      647 699  0.33     0.847
Dep:Rend       6      646 706  1.02     0.985
Dep:Cred       2      644 696  2.55     0.280
Dep:Entr       2      643 695  3.62     0.163
Dep:Ptotal     2      642 694  5.37     0.068 .
Dep:Pagas      2      641 693  6.26     0.044 *
Hab:Idcasa     4      643 699  3.90     0.419
Hab:Civil      4      633 689 13.79     0.008 **
Hab:Postal     4      643 699  3.90     0.419
Hab:Prof       6      638 698  8.88     0.181
Hab:Antg       2      645 697  1.98     0.371
Hab:Rend       6      644 704  3.05     0.802
Hab:Cred       2      646 698  0.69     0.709
Hab:Entr       2      645 697  2.34     0.310
Hab:Ptotal     2      645 697  1.62     0.444
Hab:Pagas      2      643 695  3.66     0.160
Idcasa:Civil   4      643 699  4.14     0.387
Idcasa:Postal  4      643 699  3.55     0.470
Idcasa:Prof    6      640 700  6.47     0.372
Idcasa:Antg    2      646 698  0.93     0.630
Idcasa:Rend    6      644 704  3.00     0.809
Idcasa:Cred    2      646 698  0.87     0.647
Idcasa:Entr    2      645 697  1.57     0.456
Idcasa:Ptotal  2      647 699  0.29     0.866
Idcasa:Pagas   2      646 698  0.85     0.654
Civil:Postal   4      636 692 10.72     0.030 *
Civil:Prof     6      641 701  5.66     0.463
Civil:Antg     2      646 698  0.82     0.664
Civil:Rend     6      645 705  2.29     0.891
Civil:Cred     2      645 697  1.40     0.497
Civil:Entr     2      644 696  3.10     0.212
Civil:Ptotal   2      644 696  3.08     0.214
Civil:Pagas    2      646 698  0.95     0.622
Postal:Prof    6      645 705  1.75     0.942
Postal:Antg    2      645 697  1.87     0.393
Postal:Rend    6      643 703  4.23     0.645
Postal:Cred    2      647 699  0.34     0.842
Postal:Entr    2      646 698  0.59     0.743
Postal:Ptotal  2      646 698  0.89     0.641
Postal:Pagas   2      643 695  4.24     0.120
Prof:Antg      3      645 699  1.84     0.607
Prof:Rend      9      644 710  2.44     0.982
Prof:Cred      3      641 695  5.83     0.120
Prof:Entr      3      646 700  0.85     0.837
Prof:Ptotal    3      646 700  1.23     0.745
Prof:Pagas     3      645 699  1.57     0.667
Antg:Rend      3      646 700  1.07     0.786
Antg:Cred      1      646 696  0.77     0.381
Antg:Entr      1      646 696  0.42     0.519
Antg:Ptotal    1      643 693  4.06     0.044 *
Antg:Pagas     1      642 692  4.60     0.032 *
Rend:Cred      3      644 698  3.28     0.350
Rend:Entr      3      647 701  0.05     0.997
Rend:Ptotal    3      647 701  0.01     1.000
Rend:Pagas     3      645 699  1.89     0.595
Cred:Entr      1      647 697  0.22     0.635
Cred:Ptotal    1      641 691  5.51     0.019 *
Cred:Pagas     1      639 689  7.88     0.005 **
Entr:Ptotal    1      645 695  1.94     0.163
Entr:Pagas     1      646 696  1.18     0.278
```

```
Ptotal:Pagas   1      631 681 16.24 5.6e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➔Testar sugestão do output anterior

*> ModAIC3 (Ptotal * Pagas)*

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.82e+00  -6.30e-03  -3.21e-04  -1.09e-05   2.57e+00

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.84e+01   6.32e+02   -0.06   0.9515
IdadeIdosos  -2.27e+00   8.64e-01   -2.62   0.0087 **
IdadeJovens   3.88e+00   3.17e-01   12.23  < 2e-16 ***
DepMuitos     1.24e+00   6.93e-01    1.79   0.0741 .
DepPoucos     2.78e+00   3.49e-01    7.96  1.7e-15 ***
CivilDiv      3.24e+00   5.98e-01    5.43  5.8e-08 ***
CivilSol      1.57e+00   2.28e-01    6.90  5.0e-12 ***
PostalZona3   3.39e+00   6.26e-01    5.41  6.2e-08 ***
PostalZona4   5.55e+00   6.31e-01    8.79  < 2e-16 ***
RendBaixo     1.68e+01   6.32e+02    0.03   0.9788
RendMedio     8.93e+00   6.32e+02    0.01   0.9887
RendMinimo    2.32e+01   6.32e+02    0.04   0.9707
Cred         -1.02e-05   1.64e-06   -6.21  5.5e-10 ***
EntrMenos10   5.07e+00   3.68e-01   13.77  < 2e-16 ***
Ptotal        2.46e-02   3.39e-03    7.24  4.5e-13 ***
Pagas        -1.04e-01   1.34e-02   -7.80  6.0e-15 ***
Ptotal:Pagas  1.07e-04   2.70e-05    3.97  7.1e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  639.69  on 9983  degrees of freedom
AIC: 673.7

Number of Fisher Scoring iterations: 19
```

> anova(ModAIC2, ModAIC3, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Cred + Entr + Ptotal +
    Pagas
Model 2: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal * Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9987       1487
2      9983        640  4      848    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> R2_AIC3

```
[1] 0.778
```

➔2ª função

> ModAIC3

```
Single term additions

Model:
Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal * Pagas
              Df Deviance AIC  LRT Pr(Chi)
<none>                640 674
Idade:Dep      3      639 679 1.11  0.7743
Idade:Civil    4      639 681 0.84  0.9334
Idade:Postal   4      637 679 2.61  0.6248
Idade:Rend     6      640 686 0.12  1.0000
Idade:Cred     2      634 672 5.78  0.0557 .
Idade:Entr     2      639 677 0.44  0.8041
Idade:Ptotal   2      639 677 0.31  0.8543
Idade:Pagas    2      639 677 0.80  0.6714
Dep:Civil      4      636 678 3.21  0.5230
Dep:Postal     4      638 680 2.03  0.7297
Dep:Rend       6      639 685 0.71  0.9943
Dep:Cred       2      638 676 2.05  0.3584
Dep:Entr       2      637 675 2.92  0.2320
Dep:Ptotal     2      635 673 4.70  0.0955 .
Dep:Pagas      2      635 673 4.90  0.0862 .
Civil:Postal   4      630 672 9.74  0.0451 *
Civil:Rend     6      636 682 3.26  0.7755
Civil:Cred     2      638 676 1.65  0.4372
Civil:Entr     2      638 676 1.44  0.4866
Civil:Ptotal   2      635 673 4.52  0.1045
Civil:Pagas    2      638 676 2.11  0.3483
Postal:Rend    6      636 682 3.61  0.7297
Postal:Cred    2      639 677 0.32  0.8513
Postal:Entr    2      639 677 0.46  0.7945
Postal:Ptotal  2      639 677 0.77  0.6801
Postal:Pagas   2      636 674 3.44  0.1789
Rend:Cred      3      636 676 3.65  0.3016
Rend:Entr      3      640 680 0.03  0.9986
Rend:Ptotal    3      640 680 0.02  0.9990
Rend:Pagas     3      638 678 1.71  0.6349
Cred:Entr      1      639 675 0.65  0.4202
Cred:Ptotal    1      636 672 3.56  0.0591 .
Cred:Pagas     1      632 668 7.51  0.0061 **
Entr:Ptotal    1      638 674 1.31  0.2526
Entr:Pagas     1      637 673 2.41  0.1204
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➔Testar sugestão do output anterior

*> ModAIC4 <-(Ptotal * Pagas + Cred * Pagas)*

```
Deviance Residuals:
      Min        1Q     Median         3Q        Max
-2.86e+00  -5.90e-03  -2.74e-04  -8.08e-06   2.66e+00

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.09e+01   6.29e+02   -0.06   0.9482
IdadeIdosos  -2.18e+00   8.53e-01   -2.56   0.0105 *
IdadeJovens   3.96e+00   3.23e-01   12.24  < 2e-16 ***
DepMuitos     1.26e+00   6.95e-01    1.81   0.0697 .
DepPoucos     2.84e+00   3.55e-01    8.00  1.2e-15 ***
CivilDiv      3.31e+00   5.96e-01    5.56  2.7e-08 ***
```

```
CivilSol      1.60e+00   2.30e-01    6.97  3.3e-12 ***
PostalZona3   3.43e+00   6.29e-01    5.45  5.0e-08 ***
PostalZona4   5.62e+00   6.35e-01    8.85  < 2e-16 ***
RendBaixo     1.69e+01   6.29e+02    0.03   0.9785
RendMedio     8.98e+00   6.29e+02    0.01   0.9886
RendMinimo    2.34e+01   6.29e+02    0.04   0.9703
Cred          7.67e-07   4.30e-06    0.18   0.8583
EntrMenos10   5.15e+00   3.74e-01   13.77  < 2e-16 ***
Ptotal        2.57e-02   3.36e-03    7.66  1.8e-14 ***
Pagas        -8.66e-02   1.44e-02   -6.03  1.7e-09 ***
Ptotal:Pagas  1.01e-04   2.59e-05    3.90  9.8e-05 ***
Cred:Pagas   -1.06e-07   3.93e-08   -2.69   0.0072 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2880.12  on 9999  degrees of freedom
Residual deviance:  632.18  on 9982  degrees of freedom
AIC: 668.2

Number of Fisher Scoring iterations: 19
```

>anova(ModAIC,ModAIC4,test="Chisq")

```
Analysis of Deviance Table

Model 1: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal + Pagas
Model 2: Default ~ Idade + Dep + Civil + Postal + Rend + Cred + Entr +
    Ptotal * Pagas + Cred * Pagas
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      9984        655
2      9982        632  2     22.7   1.2e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

>R2_AIC4

```
[1] 0.78
```

> ModAIC

```
Coefficients:
(Intercept)  IdadeIdosos  IdadeJovens    DepMuitos    DepPoucos     CivilDiv
  -4.37e+01    -2.15e+00     3.85e+00     1.27e+00     2.77e+00     3.22e+00
   CivilSol  PostalZona3  PostalZona4    RendBaixo    RendMedio   RendMinimo
   1.56e+00     3.45e+00     5.61e+00     1.71e+01     8.76e+00     2.35e+01
       Cred  EntrMenos10       Ptotal        Pagas
  -1.02e-05     5.07e+00     3.60e-02    -5.57e-02

Degrees of Freedom: 9999 Total (i.e. Null);  9984 Residual
Null Deviance:          2880
Residual Deviance: 655         AIC: 687
```

»»»

Glossário

A

Alavancagem (*Leverage*) significa investir mais do que a capacidade financeira permite. Designa o uso de vários instrumentos financeiros, tais como empréstimos de capital, com o intuito de aumentar o retorno potencial das operações financeiras, aumentado consequentemente também o seu risco. A alavancagem tem, portanto, um efeito ampliador tanto dos ganhos como das perdas.

Arbitragem no mercado financeiro e em economia, entende-se por uma operação de compra e venda de valores negociáveis, realizada com o objectivo de ganhos económicos sobre a diferença de preços existente, para um mesmo activo, entre dois mercados. Trata-se de uma operação sem risco (ou de risco reduzido) em que aquele que realiza a arbitragem aproveita o espaço de tempo existente entre a compra e a venda, em mercados diferentes (em que o preço do activo ainda não se ajustou) para auferir lucro.

B

Bonds ver Obrigação.

C

Capital Económico capital necessário para o banco permanecer solvente em caso de perda extrema. O capital económico é o número de desvios-padrão à distância da perda esperada que é necessário para proteger o banco da insolvência, em caso de perdas extremas na carteira do banco, devido ao risco de incumprimento. Também conhecido como "capital de risco".

Carteira de Empréstimos colecção de activos sem liquidez, empréstimos, por exemplo, mantida por uma instituição bancária.

***Cash-flow* (Fluxo de Caixa)** refere-se ao montante de caixa recebido e gasto por uma empresa durante um período de tempo definido. Refere-se à disponibilidade financeira da empresa. É a soma dos lucros acrescida da depreciação, por outras palavras, é o fluxo de dinheiro disponível. O seu aumento ou diminuição mostra a capacidade para gerar riqueza, e calcula-se com base no lucro contábil de uma empresa.

Cliente Cumpridor todo o cliente capaz de cumprir as condições de um contracto de dívida do qual é titular.

Cliente Incumpridor todo o cliente que , antes do final do contracto, falha algum pagamento.

Contraparte uma contraparte é um participante num negócio. Um negócio tem várias partes (participantes) que efectuam o negócio entre si, sendo cada um deles contraparte de todos os outros. Por exemplo, numa transacção de acções, o vendedor é contraparte do comprador, e o comprador, contraparte do vendedor.

Corporate Bonds títulos de dívida emitidos por uma empresa e vendidos a investidores, para financiar investimentos de capital e gerir o *cash-flow*. Em alguns casos, os activos da empresa podem ser usados como garantia para as obrigações. São considerados títulos de maior risco do que os títulos do governo. Como resultado, as taxas de juros são quase sempre mais elevadas, mesmo para as empresas com a melhor qualidade de crédito. A maioria das *corporate bonds* são tributáveis com os termos a mais de um ano. A dívida de *corporate* que vence em menos de um ano é geralmente designada por "papel comercial".

Crédito significa confiança/segurança. É a capacidade de assumir compromissos, quer para financiamentos, quer para empréstimos, junto do sistema financeiro. A cedência de crédito é uma relação contratual mediante a qual uma das partes, a credora, entrega um determinado montante a outra parte, a devedora, sob a promessa de pagamento do valor em dívida. O credor é toda a pessoa titular de um crédito, ou, que tem a receber de outrem uma certa importância em dinheiro.

Concentração de Risco de Crédito *(Concentration Risk)* se uma carteira de empréstimos é altamente concentrado, em termos de exposição a um sector específico ou a um país, é altamente susceptível à correlação entre ocorrência de *default* e eventos de migração de crédito. Também se designa Risco de Concentração.

CreditMetrics modelo utilizado pelo JP Morgan para a quantificação do risco de crédito em carteiras de produtos de crédito, instrumentos de renda fixa e instrumentos financeiros sujeitos ao incumprimento da contraparte.

Curva de Juros (*Yield Curve*) representação gráfica da estrutura a termo (*term structure*) das taxas de juro. Ilustra, geralmente, o rendimento local dos títulos com vencimentos diferentes mas com os mesmos factores de risco (tais como a solvabilidade do emitente), relativamente à maturidade.

D

Default falência ou incumprimento; falha no pagamento de um cupão ou do reembolso da dívida.

Default Loss para cada devedor, é o montante perdido pelo credor devido à ocorrência de default.

Derivado um instrumento derivado é um produto cujo valor varia com as mudanças que ocorrem numa ou mais variáveis do mercado subjacente, como capital próprio ou preços de *commodities*, taxas de juros ou taxas de câmbio. Derivados básicos incluem *forwards*, futuros, *swaps*, opções, *warrants* e obrigações convertíveis.

E

Empréstimo Contrato no qual uma parte fornece dinheiro, por um determinado período, a outra parte em troca de pagamento regular de juros e do nominal.

Estrutura a Termo (*Term Structure*) relação entre taxas de juro de títulos com maturidades diferentes, geralmente, representada na forma de um gráfico, designado por curva de juros (*Yield Curve*).

Exposição Vulnerabilidade à perda devido a acontecimentos imprevisíveis.

F

Frequência de Incumprimentos Esperados (*Expected Default Frequency*) a probabilidade de um devedor se revelar incumpridor antes do vencimento da obrigação. Estas probabilidades podem ser estimadas a partir de dados históricos ou analiticamente, usando a teoria de opções.

H

Hedging ou cobertura de risco, é um processo onde um risco ou exposição existente é reduzido pela adição de uma outra posição. O risco pode provir de risco cambial, de risco de taxa de juro, da detenção de acções, de risco de contraparte, etc. A redução do risco pode ser obtida através de posições não correlacionadas ou inversamente correlacionadas, pelo uso de derivados sobre os próprios activos com risco ou sobre activos correlacionados (se o derivado ganhar valor com a perda de valor do activo) ou inversamente correlacionados (se o derivado ganhar valor com a subida desses activos), etc.

M

Matriz de Transição matriz de probabilidades de transição ou de migração de uma classificação de crédito para outra.

Modelo uma representação simplificada de algo real. Os modelos facilitam a compreensão de um fenómeno e, eventualmente, a sua exploração. Algoritmos, fórmulas, sistemas ou regras que visam representar processos ou atributos reais relacionados com risco de crédito podem ser considerados modelos de risco de crédito.

Modelo de Risco risco decorrente da utilização indevida de modelos financeiros, incluindo a ausência de calibração dos modelos e especificação incorrecta.

Modelos de Risco de Crédito modelos que visam quantificar o risco de crédito - por exemplo, modelos para a recuperação de recompensa prometida; modelos de risco de *default*.

Movimento Browniano tipo de processo de Markov, que é usado para especificar o Processo Estocástico dos preços dos activos - também conhecido como o processo de Wiener.

Movimento Browniano Geométrico modelo frequentemente utilizado no processo de difusão seguido pelos preços dos activos. Movimento Browniano standard é um processo de passeio aleatório com incrementos Gaussianos, isto é, as variações nos preços dos activos são normalmente distribuídos. O termo geométrico significa que a variação dos retornos logarítmicos (em oposição à variação dos preços) é normalmente distribuída.

O

Obrigação em finanças, é um título de dívida, em que o emitente deve aos titulares da dívida e, dependendo dos termos do título, é obrigado a pagar juros (os cupões) e/ou a restituir o montante numa data posterior, denominada maturidade. O detentor de uma obrigação designa-se por obrigacionista e torna-se um credor da instituição emitente. Podem ser emitidas por um estado, governo regional, município ou empresa. Considerando-se uma obrigação como um empréstimo: o emissor é o mutuário (devedor), o titular é o financiador (credor) e, o cupão (*coupon*) é o juro.

Obrigação cupão-zero ou Obrigação sem cupão ver *Zero Coupon Bond*.

Opção contrato de que dá ao comprador o direito, mas não a obrigação, de comprar ou vender um activo subjacente a um determinado preço (preço de exercício ou *strike price*) em ou antes de uma data acordada (o período de exercício). Por esse direito, o comprador paga um prémio ao vendedor. O vendedor de uma opção tem a obrigação de comprar ou vender ao preço de exercício se o comprador exercer o seu direito.

P

Pay-Off benefício recebido ou reembolso completo de um empréstimo (principal e juros), ou a quitação integral da obrigação, ou o retorno de um acordo, decisão ou investimento.

Perda Esperada (*Expected Loss*) perda média que o banco pode esperar vir a incorrer num activo durante um período num horizonte especificado.

Perda Esperada de uma Carteira soma simples das perdas esperadas individuais de todos os activos de risco numa carteira.

Perda Inesperada incerteza do montante de perda, do valor da carteira, causada por condições de mercado. As perdas inesperadas são provocadas pela ocorrência de *default* e migrações de crédito inesperadas.

Perda Inesperada de uma Carteira perda que é inesperada ao nível da carteira. Ao contrário da perda esperada de uma carteira, esta perda não é igual à soma linear das perdas individuais inesperadas dos activos de risco que compõem a carteira agregada. Devido aos efeitos da diversificação, as perdas inesperadas da carteira são muito menores do que a soma das perdas individuais inesperadas.

Portfolio carteira de aplicações financeiras ou colecção de investimentos mantida por uma instituição ou indivíduo. Manter um portfolio de aplicações faz parte de uma estratégia de diversificação, com o intuito de diminuir riscos. Esses investimentos incluem frequentemente acções, que são os investimentos em empresas individuais; obrigações, que são os investimentos em dívida e que são projectados para ganhar juros e fundos mútuos.

Probabilidade de *default* Probabilidade de falhar o cumprimento do contracto.

Probabilidade de Transição probabilidade da qualidade de crédito melhorar ou piorar; quantifica a migração de crédito

Processo Estocástico processo que pode ser descrito pela evolução de algumas variáveis aleatórias sobre algum parâmetro, que pode ser discreto ou contínuo. O Movimento Browniano Geométrico é um exemplo de um Processo Estocástico parametrizado pelo tempo. Os Processos Estocásticos são usados para desenvolver modelos de apreçamento futuro de instrumentos em termos de preço e de algumas variáveis aleatórias, ou, analogamente, o valor futuro de uma taxa de juro ou da taxa de câmbio.

Propriedade Martingala postula que a expectativa condicional do valor futuro de uma variável aleatória é o valor actual.

Q

Qualidade de Crédito , é medido, em determinado momento, pela probabilidade da capacidade de pagamento das obrigações contratuais. É definida como *default* ou incumprimento.

R

Rating avaliação da qualidade de investimento de um emitente por uma agência de rating.

Risco de Crédito a possibilidade da contraparte falhar um pagamento. No instante em que essa situação ocorre diz-se que o devedor originou um incumprimento (*default*).

Risco de Mercado*(Market risk)* exposição ao risco de variação do valor de algumas variáveis do mercado, tais como taxas de juro ou alteração de taxas de câmbio ou preços das *commodities*.

Risco de Taxa de Juro risco de contrair empréstimos ou investimentos decorrentes de alterações na taxa de juro.

Risco Neuro *(Neutral Risk)* condição teórica em que os investidores não exigem nenhuma compensação para a tomada de risco.

S

***Spread* de crédito** taxa de juro entre duas emissões de dívida com maturidade e duração semelhantes. Reflecte a qualidade de crédito relativa dos emitentes. O spread de crédito é frequentemente usado como uma medida de solvabilidade. Uma redução no spread do crédito reflecte uma melhoria na qualidade de crédito do mutuário.

T

Taxa de Juro do Cupão (*Coupon interest rate*) Pagamentos de uma obrigação

Taxa de Recuperação percentagem do valor até ao qual o valor nominal de uma obrigação pode ser recuperado, uma vez que o devedor está em incumprimento.

Taxa Livre de Risco (Risk-free rate) uma taxa de interesse teórico em que um investimento pode ganhar juros, sem incorrer nenhum risco.

Term Structure ver estrutura a termo.

V

***Value-at-Risk* (VaR)** limite probabilístico das perdas de mercado, durante um determinado período de tempo (período de exploração), expresso em termos de um determinado grau de certeza (o intervalo de confiança). O VaR é a pior perda esperada, ao longo do período de exploração, dentro das probabilidades definidas pelo intervalo de confiança. Perdas maiores são possíveis, mas com uma probabilidade baixa.

Volatilidade medida de variabilidade (mas não da direcção) dos preços ou das taxas de juro.

Y

Yield Curve ver Curva de Juros.

Z

Zero Coupon Bond tipo de obrigação que não paga jugos periódicos; juro está implícito a diferença entre o preço de emissão (ou de aquisição) e o valor de reembolso. Tal significa que estas obrigações são emitidas a desconto.

Bibliografia

[Andrade, 2004] Andrade, F. (2004). *Desenvolvimento de Modelo de Risco de Portfólio para Carteiras de Crédito a Pessoas Físicas*. Dissertação de mestrado, Escola de Administração de Empresas de São Paulo, Fundação Getúlio Vargas, São Paulo.

[Araújo, 2006] Araújo, E. (2006). *Modelagem de risco de crédito: Aplicação de modelos Credit Scoring no Fundo Rotativo de Ação da Cidadania-Cred Cidadania*. Dissertação de mestrado, Universidade Federal de Pernambuco.

[Arvantis e Gregory, 2001] Arvantis, A. e Gregory, J. (2001). *Credit*. Risk Books, London.

[Barreto, 1996] Barreto, I. (1996). *Manual de Finanças - A moderna teoria de A a Z*. Printer Portuguesa, Mem Martins.

[Bedrick e Hill, 1990] Bedrick, E. e Hill, J. (December 1990). Outlier tests for logistic regression: A conditional approach. *Biometrika Trust*, 77(4).

[Bessis, 1998] Bessis, J. (1998). *Risk Management in Banking*. Wiley, England.

[Blatt, 1999] Blatt, A. (1999). *Avaliação De Risco E Decisão De Crédito: Um Enfoque Prático*. Livraria Nobel Sa, São Paulo.

[Cañizares et al., 2004] Cañizares, L., Montoya, A., e Zuriaga, I. (2004). *Glosario de Términos Bursátiles Y Financeiros*. Civitas Ediciones, SL, Madrid.

[Caouette et al., 1998] Caouette, J., Altman, E., e Narayanan, P. (1998). *Managing Credit Risk: The Next Great Financial Challenge*. Wiley, Hardcover.

[Chaia, 2003] Chaia, A. (2003). *Modelos de Gestão do risco de Crédito e sua Aplicabilidade no Mercado Brasileiro*. Dissertação de mestrado, Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo.

[Crawley, 2007] Crawley, M. (2007). *The R Book*. Wiley, England.

[Duffy, 1990] Duffy, D. (June 1990). on continuity-corrected residuals in logistic regression. *Biometrika Trust*, 77(2).

[Filho et al.,] Filho, E. F., Carvalho, A. C. ., e Matias, A. B. Utilização de redes neurais artificiais na análise de risco de crédito a pessoas físicas.

[Guerreiro e Mexia, 2008] Guerreiro, G. e Mexia, J. (2008). *Stochastic vortices in periodically reclassified populations*. Discussiones Mathematicae, Probability and Statistics.

[Guerreiro, 2010] Guerreiro, P. (19 Janeiro de 2010). Desempregados, divorciados e hipotecados.

[Gupton et al., 1997] Gupton, G. M., Finger, C., e Bhatia, M. (Abril/1997). Creditmetrics technical document.

[Hair et al., 2009] Hair, J., Black, W., e Anderson, R. (2009). *Multivariate Data Analysis*. Wiley, Hardcover.

[INE, 2002] INE (2002). Censos 2001.

[Jennings, 1986] Jennings, D. (Dec. 1986). Oultiers and residual distributions in logistic regression. *Journal of the Americal Statistical Association*, 81(396).

[Jouini et al., 2001] Jouini, E., Cvitanic, J., e Musiela, M. (2001). *Option Princing, Interest Rates and Risk Management*. Cambridge University, Cambridge.

[Lewis, 1992] Lewis, E. (1992). *An Introduction to Credit scoring*. An Introduction to Credit scoring, California.

[McCullagh e Nelder, 1989] McCullagh, P. e Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, USA.

[McNeil et al., 2005] McNeil, A., Frey, R., e Embrechts, P. (2005). *Quantitative Risk Management*. Princeton University Press, USA.

[Morgan, 1997] Morgan, J. (Abril/1997). Creditmetrics technical document.

[Murteira et al., 2002] Murteira, B., Ribeiro, C., Silva e, J., e Pimenta, C. (2002). *Introdução à Estatística*. MCc Graw Hill, Portugal.

[Natário, 2009] Natário, I. (30 Julho de 2009). Probabilidades e estatística d.

[Nelder e Wedderburn, 1972] Nelder, J. e Wedderburn, R. (Dec. 1972). Generalized linear models. *Journal of the Royal Statistical Society*, A(135).

[Neto, 2002] Neto, A. (2002). *Modelagem de risco de crédito: um estudo do segmento de pessoas físicas em um banco de varejo*. Dissertação de mestrado, Universidade Federal de Pernambuco.

[Oksendal, 2007] Oksendal, B. (2007). *Stochastic Differential Equations*. Springer.

[Pereira, 2009] Pereira, J. (2009). Credit risk.

[Ribeiro et al., 2006] Ribeiro, E. M., Neto, J. D., Melro, E. M., e Mello, C. R. (2006). Aplicação das redes neurais na concessão de crédito - um estudo de um caso em uma empresa de consórcio.

[Rutkowski e Musiela, 1998] Rutkowski, M. e Musiela, M. (1998). *Martingale Methods in Financial Modelling*. Springer.

[Saunders e Allen, 2002] Saunders, A. e Allen, L. (2002). *Credit Risk Measurement: New Approaches To Value-At-risk And Other Paradigms*. John Wiley, Hardcover.

[Schrickel, 2000] Schrickel, W. (2000). *Análise de crédito: concessão e gerência de empréstimos*. Atlas, São Paulo.

[Securato, 2002] Securato, J. (2002). *Crédito: Análise e Avaliação do Risco: Pessoas Físicas e Jurídicas*. An Introduction to Credit scoring, São Paulo.

[Silva, 2000] Silva, J. (2000). *Gestão e Análise do Risco de Crédito*. Atlas, São Paulo.

[Turkman e Silva, 2000] Turkman, M. e Silva, J. (2000). Modelos lineares generalizados-da teoria à prática.

[Vasconcellos, 2002] Vasconcellos, M. (2002). *Proposta de Método para análise de concessão de crédito a pessoas físicas*. Dissertação de mestrado, Faculdade de Economia,Administração e Contabilidade, Universidade de São Paulo.

[Wang e Salous, 2009] Wang, Z. e Salous, S. (Fev. 2009). Spectrum occupancy statistics and time series models for cognitive radio. *Journal of Signal Processing Systems*.

[Wikipédia,] Wikipédia. Vasicek model.

[Wikipédia, 2009a] Wikipédia (2009a). Akaike information criterion.

[Wikipédia, 2009b] Wikipédia (2009b). Análise de crédito.

[Wikipédia, 2009c] Wikipédia (2009c). Regressão logística.

[Wikipédia, 2009d] Wikipédia (2009d). Z-score financial analysis tool.

[Zagst, 2002] Zagst, R. (2002). *Interest Rate Management*. Springer, Germany.

[Zerbini, 2000] Zerbini, M. (2000). *Três ensaios sobre crédito*. Tese doutoramento, Faculdade de Economia,Administração e Contabilidade, Universidade de São Paulo.